现实挖掘

[美] 内森·伊格尔 凯特·格林◎著
吕荟 陈青青◎译

现实挖掘是即将改变世界的十大技术之一。——《麻省理工学院科技评论》

数据挖掘不可不知的采集要领

大数据商业化的必经之路

摆脱数据困境，高效利用数据价值的关键

REALITY MINING

Using Big Data to Engineer a Better World

Nathan Eagle Kate Greene



中信出版集团·CHINACITICPRESS

版权信息

书名:现实挖掘

作者:内森·伊格尔 凯特·格林

译者:吕荟 陈菁菁

ISBN:9787508656434

中信出版集团制作发行

序言

大数据正在席卷全球。这个话题如今频繁见诸各种会议、专著、论文和企业的讨论中。这当然是事出有因的：对以往深不可测的大量数据进行挖掘，从而发现趋势甚至预测未来，这样的想法的确非常具有吸引力。但是正如这些会议、专著、论文和商业计划中所阐述的，弄明白如何处理巨大体量的数据，并使其发挥更好的作用并不是一件简单的事情。

我们将大数据定义为人或物与数字网络世界之间相互作用而产生的信息集合。它可以是数年间采集的关于一个人的单一变量的数据，也可以是在某一瞬间采集的关于数亿人的多变量数据。大数据可能涉及的时间长、囊括的主题多或者涵盖的范畴广，也可能是这三种特征间的组合。

由于各种技术因素的汇集，大数据已经成为我们现代世界的一个特征。高性能的移动互联智能设备就在你的口袋中采集数据、进行运算，然后发送给远程服务器。云计算和日益增长的高密度数据存储设备，为一切信息提供了安身之所。并且，流处理范式使数据可以通过分布式设备进行处理。一些针对大规模数据集设计的编程模型，如MapReduce以及开源的Hadhoop的出现，使人们了解即将到来的信息流是可能的。

大数据被定义为我们日常活动产出的数字记录或留下的数字足迹，它是我们生活的元数据。一些人害怕这会带来一个没有隐私的世界：企业对我们的了解比我们自己还多，政府可以监控那些它们认为危险的人。另一些人则认为大数据是数据库这抹彩虹末端的一罐金

子，是抓住下一波信息技术趋势的机遇。他们还相信，从这些全世界人们日常生活中产生的海量数据中可以获得有益的东西。

作为本书作者，我们是技术专家，属于后者。我们认为，如果从负责任的、审慎的以及对环境敏感的角度来看，大数据可以帮助改善公共卫生、引导个人更好地决策、促进知识的共享以及提升创新速度。大数据的时代已经来临，而且看起来也不会很快结束。因此，保证个人自由和隐私不被侵犯，告知消费者谁在什么时间、以什么为目的获得了他们的数据就很重要。我们相信，在小心谨慎进行数据采集的前提下，大数据就可以被用来设计成更好的系统，以及一个可能的更好的世界。我们采用了一个被称为“现实挖掘”（Reality Mining）的方式，不仅对大数据进行分析，而且确保分析能够反映参与人的现实状况，同时在整个过程中始终保持谨慎的数据采集态度。

本书的目标是探讨大数据可能的积极方面，特别是展现如何利用对现实的挖掘来设计更好的社会系统。这意味着本书所介绍的理念，将会超越那些简单的描述性分析，如计步数据的柱状图等。我们所探讨的是针对一些可视化的数据表达，比如犯罪行为或疾病暴发的空间分布图等，如何将其转化为具有实际操作意义的行动或政策。我们还考虑了可以使用全球的匿名数据系统的构想。提出诸如“如何在缺乏公共卫生资源的发展中国家，配置疾病传染的预警系统？”这样的问题。根本上，我们希望探寻如何利用大数据让人们的生活变得更加美好。

本书分为5个部分，每个部分分别关注不同的数据采集尺度，大数据的内在挑战和机遇。大致同查尔斯·伊姆斯和蕾·伊姆斯夫妇在1977年制作的电影短片《十的力量》中分别观察了宇宙的极大部分和极小部分一样，本书对大数据的讨论也是从小样本渐增到较大的样本。

本书的现实挖掘旅程将从个体层面开始，即单个人所产生的数据和应用于单个人的数据。接着，我们升级到邻里和组织层面，再扩展到城市层面，之后是国家层面，最后到达全球层面。诚然，这样的分

层框架并不是绝对的，也无法囊括所有的场景类型。当然，在某一层面获取的大数据也可以应用于其他不同层面。我们适时地提及了一些这类多样化应用的例子。不过，5个层次的结构更易把握，也更有助于我们思考在大数据获取和利用过程中面临的种种挑战。

这5个部分又分别各有两章。每个部分的第一章详细描述了该层面的数据采集种类、数据采集方式，以及读者在可能的情况下可以如何获得这些数据；第二章则阐释在这些数据的基础上，已经建立或者我们相信可以建立的应用和系统。

因此，每部分的第一章作为向导，带领读者对可以挖掘的多种数据类型以及可能的获取途径进行思考。这些途径可能是编写一个手机应用程序来采集使用者的睡眠数据，或是订阅服务商提供的航班数据，又或是基于谷歌进行检索的数据分析。有些多数人不太容易获取的数据，比如移动电话的通话记录，我们也提供了如何有限获取或者寻找其他可能来替代实际数据的建议。

在充分理解哪些数据类型可以被挖掘之后，我们接着讨论可以使用这些数据进行现实挖掘的程序。有些系统较为成熟，一些则还处于早期阶段，其他则尚未被开发出来。我们在本书中仅提供了一小部分可用程序的信息，同时也展现了机遇之所在。

隐私问题是大数据采集和使用过程中的大麻烦，在本书写作过程中，我们曾考虑用单独的一章来讨论它，但最终放弃了这个想法。工程师和企业往往在最初的产品构想已经基本完成时，才开始考虑用户或客户的隐私问题。这样的话，隐私特性便成为几近完成的主要设计的添头。我们不希望这本书也变成这样。我们认为，人们对隐私的期望和分享数据的意愿度应该从一开始就被考虑，并烙在每一个应用程序或产品设计中。故依据需要，我们通过探讨隐私问题、人们对数据采集和共享方式的了解程度、人们对这些方式的适应度（适应度往往

受到多种因素的影响而不断变动）以及在注重隐私的前提下开发应用来反映这种情绪。

需要特别指出的是，本书没有涉及特定的分析方法论，而是将这些留给了其他文章、论文和讨论。大数据技术还在不断发展，现实挖掘的实践者们很快会发现将有更多分析技术可用于他们的数据集。在本书中我们没有排斥某些特定的数据和应用程序的分析，而是选择聚焦于更宽泛的现实挖掘问题：如何安全、不侵犯隐私而又有意义地进行数据采集？如何设计实用、以人为本的系统？

关于大数据的很多讨论都围绕挖掘“知识”这一主题，似乎“知识”就是人们唯一能够且应该从中获得的。本书从另一个视角来看待大数据，在描述性分析的基础上前进了一大步，从知识走向行动。“现实挖掘”是为了使用大数据来开发系统，从而对从个体到全球的所有层面都产生积极影响，它将提升我们的生活品质，让我们变得更健康，并让我们与70亿地球邻居们一起生活得更加美好、智能和幸福。

REALITY
MINING

Using Big Data
to Engineer a Better World

第一部分
从个体开始挖掘

手机一旦掌握了你的使用习惯，它就可以帮助你安排行程、提供活动建议，或者在没有手动设置的情况下发出提醒。它可以调整使用模式以适应多样的环境，比如在影院时自动关闭手机铃声，电影结束后再自动打开。熟知你生活习惯的手机还可以为你推荐酒吧，那里的常客们跟你有着相似的爱好，或者刚好在你想要尝试去一家新餐厅吃晚饭时向你推荐一个正合你意的餐馆。

REALITY

MINING

Using Big Data

to engineer a Better World

第一章
智能手机、传感器和生命记录

如今，采集我们自身的众多日常数据已相当容易，各种各样的技术通过移动电话、软件、皮肤电流监控器、可穿戴摄像头等，追踪着我们的习惯、位置、消费、路线、社交以及情绪。的确，因我们日常行为而产生的“数据排放”，其释放和捕捉的便利性给研究人员们带来了新的机遇。不仅使研究者可以更深入地了解这些行为，还有助于他们利用这些观察来设计更适应人们实际行为的系统。

传感器、软件以及它们在我们生活中的广泛存在是推动这一趋势的重要因素，而一类重要的传感器就植入在如今的移动电话中。随着移动电话的高度普及，它成为全球几乎所有人的必需品，它已经当仁

不让地成为采集个体数据的最基本工具。到2012年年底，全球已有近60亿个移动电话客户端。即使是最简单的手机，每次与通信基站交换信号时，也会提供其使用者的位置信息。移动电话最初只是通信工具，如今已逐渐成为装配了多种附加传感器的随身计算设备。这些附加装置包括可以监测身体活动的加速仪、可以测度我们位置的GPS（全球定位系统）芯片、蓝牙模块等近距离无线传输协议、可以推断附近情况的麦克风等，甚至简单的通话记录也可以用来衡量我们的社交进展。

手机一旦掌握了你的使用习惯，它就可以帮助你安排行程、提供活动建议，或者在没有手动设置的情况下发出提醒。它可以调整使用模式以适应多样的环境，比如在影院时自动关闭手机铃声，电影结束后再自动打开。熟知你生活习惯的手机还可以为你推荐酒吧，那里的常客们跟你有着相似的爱好，或者刚好在你想要尝试去一家新餐厅吃晚饭时向你推荐一个正合你意的餐馆。

移动电话中的数据还能够提供人们的实时位置移动信息，在研究疟疾或流感这类疾病时，它可能成为建立传染路径模型的关键数据。另外，研究人员还发现，通过手机上合适的传感器和软件捕捉用户的行为变化和谈话模式，可以比其他医学检查更早发现某些疾病的预兆，如抑郁症或帕金森综合征。在使用个人数据让人们的生活变得更加轻松和健康方面，现实挖掘拥有很大的潜力，这些只是几个仍处于早期阶段的应用案例。

追踪我们个人信息的不仅仅是移动电话，我们的个人计算机使用记录也在被越来越多的软件监测。研究人员推测，人们越了解自己在某些网页或者电子邮件上花费了多少时间，就越容易掌握和调整日常效率。当然，由于移动电话变成了空前强大的计算设备，在手机上追踪人们应用程序使用记录的软件也被开发出来。将这些软件数据与通

过手机传感器和其他程序采集到的数据放在一起，可以推断出很多个人行为信息。

除了移动电话和个人计算机之外，人们在日常生活、运动或是睡觉时，也越来越多地主动携带各种专用传感设备，以掌握自己的生理习惯和健康状况。谷歌眼镜，实际上是一个装配了小型显示屏、摄像头、麦克风、处理器和无线通信的头部穿戴系统。因其通过连续拍照和摄像的方式使你与外部世界保持联系并记录你的生活而广受关注。更为普通且不显眼的计步器和睡眠监测仪正在获得商家的关注。这些设备和手机上模拟类似功能的应用程序所产生的数据，可以精确地显示一个人的身体活动状况。亲眼看到这些数据可以激励人们生活得更加健康。近年来，脸谱网（Facebook）和推特（Twitter）这些社交网站的日益流行，已经形成一个庞大的个人数据库。人们在这些站点上进行“状态更新”，发布可以反映其生活状态的短消息，回答诸如“你在做什么？”“你感觉怎么样？”“你周边发生了什么？”“现在有什么让你感兴趣的事情？”这样的问题。可以说，状态更新在某种意义上来说相当于用户对一个社会学家的社会调查问卷做出的回答。

一经发布，这些状态更新就会被推送给社交网络中的其他人，在某些情况下还是在线公开的，可以被任何想看的人看到。一些研究人员在探索根据日历事件和位置信息，自动进行状态更新的方式。另一些人则试图将这些状态信息集中解读，已有软件开发者编出一些简单程序来分析这些内容。这些程序通常将特定关键词进行可视化，依据词汇出现的频率设定字体的展示大小。通过这种方式，人们可以大致了解自己一段时间内的活动和感受。

越来越多的人开始着迷于自我的个人数据，并将他们能获得的所有自身信息结合起来，包括手机通信、电脑使用、生物传感器、摄像或手工录入数据记录等。这种极端的量化和编目方式就是所谓的“生命

记录”。尽管这种方式并不算普遍，它作为一种更好地了解自身习惯的方法，已经吸引了一些人。

工程师和设计师推动了生命记录的发展，因为他们发明了让人更容易进行生命记录的网络应用和其他技术工具。随着谷歌眼镜和其他生命记录工具的出现，更重要的是，这些工具与人们日常生活结合得更加紧密，生命记录对普通人来说将不再困难。这种活动甚至可能克服社会成见，并被更多不精通技术的人们所接纳。

在本章中，我们将探讨个人数据可以通过哪些方式被采集和记录，包括不易察觉的移动电话日常交互，以及更具目的性的电子公告，如状态更新信息等。我们还将讨论，个人和企业在数据采集和分析过程中都需要时刻铭记的隐私问题，以及目前在实践中的隐私保护方法。

麻省理工学院的数据追踪项目

2004年秋天，麻省理工学院开展了一个研究项目，向100名新生和在校学生提供预装定制版ContextPhones程序的诺基亚6600型手机。该程序可以追踪手机连接的通信基站代码、应用使用情况以及手机是否闲置或是在充电等状态信息。在9个月的时间里，该研究项目记录了30万小时的用户数据。

所有的参与者都被告知了电话记录程序的功能，并被要求签署同意书，签署意味着他们知道自己的手机采集的是什么数据。参与者可以在任何时候抹去他们自己的数据并关闭记录功能。由于和研究无关，参与者的手机号码作为附加的隐私保护措施，此次研究中将其通过单向函数（MD5）生成唯一代码，并无法被逆运算得到原始手机号码。

诺基亚6600手机本质上是一组被参与者们几乎一直随身携带的传感器。过去很多年里，大学和企业的研究人员通常使用放置在房间、办公室或是设备包中的传感器，来采集个人的位置、与其他人的位置关系、物理位移甚至是周边环境声音的片断等信息。装有红外线或RFiD（无线射频识别发生器）的智能胸卡被设计为可以识别其他同类装置，它们被用来研究工作场所的协作以及会议中的社交。这些传感器和智能标记系统相比过去的大包传感器和电路板已经有了很大的改善，但还是有些笨重。

尽管在人身上装传感器的方法有很多，但这个诺基亚6600的现实挖掘项目是独特的，因为这是有史以来第一次，研究人员以可扩展的方式追踪研究对象的位置、社会交往和习惯。项目进行时，已经有数以千万计的手机具备运行诺基亚6600上所安装的超级监测软件的性能。这个项目证明了手机可以成为可靠、可扩展的泛在计算工具，可以获取比以往任何时候都多的行为数据。在行为研究方面，过去大多数社会学家采用调研的方法所获得的信息，在规模和精确度上都无法与移动电话采集的数据相提并论。

选择诺基亚6600型手机是因为它装备了塞班S60系统软件平台，该手机可以运行由赫尔辛基大学开发的定制版Context软件，用以记录个人手机的所有状态：无论是拨打一个电话还是充电乃至闲置状态。这款手机备有6MB（兆字节）的内存，并可以使用最大32MB的多媒体闪存卡进行容量扩展。手机没有任何锁定功能，并且可以使用任何一家全球通信系统的移动电话运营商，如T–Mobile（德国电信）,AT&T（美国电话电报公司）和Cingular公司（已被AT&T收购）。定制的应用程序可以通过通用分组无线服务数据网络、蓝牙、存储卡或红外端口等方式安装到手机上。

手机可以持续不断地扫描和记录它周边一定范围内的蓝牙设备。蓝牙是一种频率在2.4兆~2.48兆赫兹之间的无线通信协议，在1994年由

爱立信公司开发，并于1998年发布，用以替代设备间的串口连接方式。每个开启蓝牙的设备都具有“设备被发现”能力，它会寻找并发现周边5~10米范围内其他蓝牙设备的媒体访问控制地址。

该研究项目使用了一个修改版本的BlueAware应用程序（移动电话版MIDP2–Java），对发现的蓝牙设备识别地址在邻近的服务器上进行记录和时间戳记[注]。因为如果使用标准版本的BlueAware软件进行持续的扫描和记录的话，手机电池会在18个小时内耗尽，所以修改版本将扫描时间改为每隔5分钟一次，从而使手机续航时间增至36个小时。当传感器装在移动电话或者其他使用电池的设备上时，考虑传感器的资源需求和维持其持续扫描的其他可行性就至关重要。

与BlueAware软件类似的一个软件是Bluedar，是为装置在研究参与者的社交聚会场所的设备而研发的。Bluedar对可发现设备进行持续扫描，并将蓝牙设备识别地址通过一个802. 11b的无线网络上传到服务器上。设备具有一个包含了第二代蓝牙芯片的核心，可以被XPort型网络服务器所控制，并能有效探测识别周边25米范围内的蓝牙设备。

除了蓝牙扫描功能，诺基亚6600还能够持续记录通信基站的识别号码。利用通信基站识别码获得用户定位数据已有大量的相关研究，但是通过这种方式获得精确的位置信息比较困难，因为手机可以连接到几英里远的通信基站上，而在城区时可连接的范围内常常会有数十个通信基站。

在这项研究中，当研究对象在某一位置停留的时间足够长，足以提供一个预估的信号塔概率密度分布函数时，那么就可以获得相对高精度的位置。由于多种条件的差异，包括信号强度和网络状况等，手机在同一位置不同时段可能分别与不同的信号塔通信。所以，在一段时间内，手机可能连接了好几个信号塔。另外，手机位置的细微改变，甚至都会引起信号塔分布的显著变化。信号塔识别地址还可以与

其他静态蓝牙设备如台式计算机相互参照，以进一步确定移动电话位置。

当然，如今手机已经可以通过很多更直接的方式记录用户位置。很多智能手机装配了GPS芯片，而谷歌和Skyhook等公司使用三角定位技术，来弥补通过无线基站信号、通信基站和GPS定位不足（在室内使用时精准度不足）的问题。但对于普通手机来说，通信基站识别依然是最经济、最简单也是最可靠的位置追踪技术。

在麻省理工学院这个基于诺基亚6600的研究早期阶段，采集的数据被存储在手机有限的内存中，研究者需要经常转存数据，这一过程需要大约5分钟，同时研究人员还可以升级应用程序。然而由于一个月的数据量就有5~10兆，一部分数据就需要被存储在手机的外置闪存卡上。通过对程序在存储卡写入数据效率的改进，数据转存的需求被延长至数月一次。到后期，使用T–Mobile作为服务商的用户还可以使用有限的互联网连接将数据通过电子邮件发送至代理服务器。

最后，研究参与者还完成了一次问卷调查，回答包括他们的手机使用情况、日常行为模式、对麻省理工学院的满意度、他们的社交圈以及工作圈等问题。问卷的最后一个问题涉及所有研究参与者，要求每个人对自己与其他每个人之间的互动频率进行打分，并确认其他人是否在自己的工作或朋友圈里。这些调查获得的信息作为手机中获取数据的补充，帮助人们更好地进行数据分析。

这个针对百人的研究项目的成果非常鼓舞人心。一个名为“本质行为”的计算机分析工具进行分析所得结果表明，通过研究某个对象在一天中早晨的所在位置、与其他参与者的空间接近程度、通话记录和通话活动等，可用于推测其当天晚上的行为。例如，某个人在一个星期六的上午10点醒来，那么可以比较准确地预测当天晚上他会跟哪些人一起出现在什么地方。另外，对通话记录及空间接近程度数据的分析，能够得知研究参与者的社交网络关系和社会地位，比如，该参与

者是学校的新生、研究生还是教授？下一章我们将讨论这项研究的成果可以有哪些进一步的应用。

自从诺基亚和麻省理工学院的这项合作研究开展以来，大学和企业实验室的很多研究项目都开始利用移动电话作为传感器来采集多种数据，尽管主题不尽相同。例如，2009年达特茅斯大学的卢弘等人的“声音感知”项目，通过iPhone（苹果手机）上安装的软件，以一种低功耗和保护隐私的方式捕捉周围的声音，判断其所处的环境。因为如果手机能够根据声音特征，判断出一个人处于重要会议中，就可以将一些人的来电直接转至语音信箱，而另一些通话则可以被接入。

麻省理工学院的媒体实验室开发了知名的Funf开放感知框架系统，作为一个开放和公开的系统，它可以被用于追踪手机的各种活动。Funf框架可以从手机上的多种“探测仪”获取数据，例如： GPS、定位仪、加速仪、通话记录、应用使用情况、屏幕状态以及电池状态等，并加密存储在手机上。在Funf这样一个基础框架上，任何一名开发者都可以进一步开发符合自身需求的软件。Funf Journal则是一个现成的安卓系统手机应用程序，它可以安全地将探测到的各种手机数据加密存储。人们还可以将数据下载到自己的电脑或上传到远程服务器上，以供进一步分析。

除了Funf框架，还有很多为手机和个人电脑设计的应用程序，用于获取商业性的数据。如果你无法基于框架自己开发软件，那么这些简单易用的商业应用程序是不错的选择。特别是，有些应用程序可以通过监控个人电脑在某些网站或是某些软件上的使用时间，从而推算计算机使用者的工作效率，这已经发展成为一小块细分的专业市场。

手机——最佳人体传感器

麻省理工学院的现实挖掘项目为通过开展以手机为基础的研究来采集个人数据这样的做法提供了良好的范例。研究中使用了赫尔辛基大学开发的一个定制软件，可以追踪手机的各种状态，如通话中、充电中、关机状态等。尽管作为一个研究项目的定制软件，一般人没有机会接触它，但是市面上还有很多其他具备类似记录功能的软件。其中一些软件可以直接录入GPS数据，而不需通过移动运营商或是使用Wi–Fi（无线网络）三角定位法来获得手机的位置信息。

在互联网上搜索一下，就能找到很多网站出售这些可以安装在苹果手机、黑莓手机，以及装有安卓、微软和塞班等手机操作系统的软件。这些软件有的是为那些担心使用手机的孩子和孩子所处位置的父母而设计，有的是为想要掌握其员工使用公司配备手机情况的雇主而设计，也有的是为怀疑另一半有不忠行为的人而设计的。需要注意的是，不同的国家和地区，对于从个人移动电话采集数据有相应的法律规定，想要合法地采集数据，需要征得手机所有者或是设备所有者的同意并签署合约。

追踪软件安装在手机上并在后台运行，时刻记录着手机的一切活动。这些记录稍后会被发送到远程服务器上，并可通过一个网站获得。包括呼入和呼出号码、通话时长以及时间戳记在内的通话记录会被采集；即使在手机上已经被删除，文字短信也会被完整地存储在远程服务器上；还有GPS位置信息也能在具备信号的地方被记录下来。有些具备特定功能的手机，其网址访问记录也会被存储下来。

针对个人计算机的类似商业监控软件也同样存在。这类软件的销售主要面向那些希望掌握自己计算机使用情况，希望借此进行自我管理并提高生产效率的人。全世界现在有数百万人每天使用计算机的工作时间都长达数小时，包括使用电子邮件、即时通信、网页浏览软件以及文字、图片和基本上所有在电脑上运行的软件如视频处理软件。

而像RescueTime、Klok、SlimTimer和WorkTime这类软件，可以追踪记录前台软件运行的时间，并将数据反馈给用户。

这些软件有的是直接安装在电脑上，有的则是通过网页运行。某些情况下，人们可以将这些软件捕捉的信息打上标签，例如读新闻故事，并分享使其成为“公共的”信息。人们还可以设定某些特定应用的目标使用时间，比如每天早晨用30分钟回复电子邮件。然后软件会以可视化的方式来展现设定目标的完成情况，并且使用者还可以设定任务变更提醒，以防自己在某一应用上耗费的时间过长。

除了在个人设备上进行自动追踪记录的后台软件外，还有一类基于调查的手机软件也在迅速发展。2009年秋天，Techneos公司发布了SODA系统，该软件系统使研究人员、调研公司在内的任何人都可以在手机上发布调查问卷。因为对大多数人来说，手机是随身携带的。与在线问卷相比，手机问卷可以选择在特定时间或特定地点进行，与电话问卷相比又不会显得那么唐突。

SODA对于想要为现实挖掘采集信息的人来说，将会是一个有趣的工具，因为相较通常意义上的问卷调查，它可以让人们提供更多的相关信息。比如经参与调查者的许可获得其位置信息，他们还可以在问卷中提交图片。这个平台非常灵活开放，问卷问题种类非常多样，可以是多项选择、浮动计算、开放数值或文本、声音文件、图片文件甚至是条形码输入。另外，该平台还支持多种语言，目前包括汉语、英语、法语、葡萄牙语、西班牙语、德语、印度语、日语和泰语。

更加精确的生物传感器

通过电脑或手机记录一个人的工作效率、通信记录、进行问卷调查等，只是人们日常生活场景中的一部分。生物传感器则可以填补这

一空缺，它可以精确记录每日不同时间和不同活动中人们的生理变化。研究人员可以借助一些特定的硬件设备追踪被研究者（通常是那些慢性疾病患者）的心率、血压、皮肤导电性以及其他一些指标，或是在一些需要手动记录症状、饮食和运动信息的情况下，也需要借助这些硬件设备。

BodyMedia是一家知名的设备以及在线服务供应商，这些设备和服务可以追踪身体活动并根据健康需要适当地发出提醒。CardioNet公司则提供便携式心电图仪器。还有诸如FitBit计步器、Nike+运动追踪系统、Polar和Garmin的GPS手表、Withings的无线体重分析仪等也都在市场上占有一席之地。（Zeo公司的个人睡眠教练同样也在市场上获得了一定的成功，它是一个追踪睡眠状况的无线头带，可以根据数据分析情况发送指令到床边的闹钟或手机上。但很不幸，这家公司于2013年解散了。）另外，越来越多的手机应用程序也开始使用手机上的加速度仪和GPS传感器来采集生物计量信息。表1–1列出了一些用于记录生物数据的商业技术。

软件开发者们也开发出了可以提供专业睡眠追踪的某些相同功能的手机应用程序，虽然精确性仍不稳定。iSleepTracker和Sleep Cycle就是其中的两个，它们利用手机的加速度传感器感知用户睡眠时在床上的动作。类似的还有安卓平台上的Smart Alarm应用，它可以预判人们的睡眠阶段。因为这些应用程序并不能直接衡量一个人的动作（而且判断的结果会受床上的人或动物数量以及床垫性能的影响），它们的主要功能似乎是叫醒处于浅睡眠阶段的人，从而避免人们在深度睡眠时被弄醒而产生不适。

表1–1 商业化感知设备的用途、性能和交互方式

产品	用途	特性	交互方式
Body Media公司的Body Media FIT	监测睡眠质量和消耗的卡路里，主要适用于有减肥需求的用户	臂带形式，其中装有一个用于监测动作的三轴加速仪，一个测量皮肤温度的温度计，一个测量皮肤水含量的皮肤电反应感应器，以及一个测量身体热量散发的热流传感器	臂带上装有简单的LED指示器；更详细的数据会通过设备定期上传到服务器上，用户可以在网站上获取数据
CardioNet公司的移动心脏门诊遥感器	为心律不齐患者监测心跳和其他问题	一个小巧的，可穿戴的心电图仪器，可以24小时监测心跳，最长可以连续记录21天	贴在佩戴者胸前的电导线将电子信号传输到一个小型的便携监控器上，这个监控器可以戴在脖子上。当心跳异常时，监控器可以通过无线传输数据到分析中心，在这里，数据会被进一步分析并给医生发送报告
FitBit公司的FitBit监测器	监测特定类型的低强度身体活动以及睡眠质量	可以夹在裤子或文胸上的一个小型的、不显眼的计步器。每次充电后可以使用10天	一个家用基站从计步器上无线传输数据到计算机服务器，并上传到专门汇总这些数据的网站
Zeo公司（该公司于2013年解散）的Zeo个人睡眠教练	监测睡眠期间大脑和脸部肌肉的活动	包括一条头带，用于测量睡眠阶段和质量对应的脑电波活动。数据通过无线传输到一个专门的闹钟上，使其在人们睡眠最浅的时候叫醒他们	一个显示睡眠信息的专门闹钟，以及一个分析睡眠信息并把睡眠质量和不同的生活方式关联起来的网页服务
Innovative Sleep Solutions公司的iSleep Tracker	监测睡眠活动和质量	一个类似手表的设备，用于监测睡眠活动	有一个提供分析和可视化服务的网站
WIN Human Recorder公司的HRS–I	监测心跳和其他问题，同时还监测体温和活动	一个小型的感应器组合，贴在使用者胸前，每次充电可以使用3~4天	数据被传输到手机或电脑上，可以在线浏览
Garmin Forerunner 910XT GPS设备	记录步行或跑步的时间、距离、海拔以及心率，还有游泳时的距离、效率、摆臂次数以及泳池的水下深度	一个类似手表的设备，监测不同模式的体力消耗和移动距离	数据通过无线传输到一个提供数据分析和分享的网站——Garmin Connect

Nike+程序利用了苹果手机、安卓手机以及iPod（苹果音乐播放器）这些人们经常在跑步时用来听音乐的设备，通过夹在鞋子上或置入在耐克某些特制鞋袜的计步设备采集跑步数据，并通过无线设备传输至手机或音乐播放器上。当设备通过用户的计算机连上互联网进行

同步更新时，数据会被上传到耐克公司的网站上，用户可以看到自己的进步，并可以与其他用户的数据进行实质比较。这一系统的目的是帮助积极跑步或徒步的人记录其锻炼成果。

还有RunKeeper、Runtastic、Runmeter等手机应用可以在手机与GPS卫星直接联系信号条件较好的情况下，通过GPS定位来追踪人们的户外活动。人们也可以手动输入一些非户外活动的信息，如在室内游泳等。某些健身设备还可以与手机直接相连并同步输出数据。所有这些信息都可以被集成到一个可视化图表中，展现个人的锻炼进展，以及不断向目标靠近的过程。

RunKeeper应用还与Withings公司建立了合作。Withings的无线体重分析仪可以在人们每次称体重时，将数据自动同步至RunKeeper网站。该数据可以被用来帮助粗略估算用户以一定速度跑完特定距离需要消耗的热量值。

上面提到的这些应用和设备都提供了多层级的用户控制功能，以便用户控制被采集数据的导出方式。有些应用自带免费的分析功能，需要通过付费以获得一些高级功能。对Body Media公司来说，研究人员就必须购买许可才能获取设备采集到的所有数据，而没有参与研究项目的个人用户，则需要购买该公司的在线服务并提供大量个人信息，才能获得。虽然这些产品为我们提供了远比任何时候都多的对于自我身体的认知，用户还是应该认真阅读产品用户协议，以确保清楚地知晓数据是如何采集和使用的，以及数据的所有权归属等问题。我们将在第二章继续探讨数据所有权的问题。

可以感知周围环境的机器学习

很多生物传感器和手机应用都具有向第三方发送数据的功能，这个第三方可以是你的医生，也可能是公众可以获取健康数据的公共网站。本质上，这是人们对自己身体状态进行的“状态更新”。近年来，公众热衷于各种各样的状态更新，因此提供了大量关于人们行为习惯的信息。

有些行为追踪应用程序可以在一个活动完成后，如一次跑步或自行车骑行后，自动更新推特或脸谱网的个人状态。前面提到的达特茅斯大学的“声音感知”项目利用手机的麦克风获取声音信息，从而推断出一个人所处的位置环境和正在从事的活动，并可以用来简单更新个人状态，如在一家咖啡馆中、在户外步行或是正在刷牙等。这个程序通过机器学习技术对捕捉的声音片段进行分析，对音乐、人声等一些基本的声音进行判断，并对用户的声音特征进行自我学习。当然，这也是在保护隐私的前提下进行的，原始声音片段并不会被存储下来，只是被处理并提取特征信息，这些特征信息并不足以被用来再现原始的声音片段。

生命全记录

很明显，采集个人的数据和信息可以有多得无法计数的方式，但是尚没有统一的方法来采集各种类型的数据，这些数据目前可能是由不同软件及设备自动获取，还有可能是人工记录的。不过已经有越来越多的人正在试图创造一种可以记录其全部或部分生活的方式：他们希望关于自己的一切都可以被量化。

生命记录作为一种趋势，在技术导向型的社会群体中更受欢迎。其中一部分人开发了在线程序和手机应用来记录数据，专门的硬件，如穿戴式摄像头来采集视频和图像，以及电子表格用以记录日常活动，如吃东西的感受以及不同的情绪等。生命记录是其他数据采集方

法的补充，并试图在一个紧密框架下整合数据对个人生活进行定量化的描述。

生命记录具有无限的潜力。有一个明显的作用是它可以让人们认识到自己的习惯，以及习惯的小小改变会如何改变他们的生活。不过其实还有更多意义更深远的可能性。如今，没有人真正了解心脏病的纵向指标究竟是什么，但是随着越来越多的人对自身生活如此多的生活片段进行自我监测并将其分享，研究人员可以利用这些数据来回顾疾病患者几个月或是几年间的数据，并确认一些相关性、可能诱因，以及严重疾病的潜在指标。

最引人注目的那些生命记录项目，往往都采用了自动影像捕捉技术。2009年年末，动作捕捉设备生产商美国威康公司授权微软公司使用其技术，用于制造和销售一款可以自动拍摄一系列贯穿全天照片的穿戴式摄像头。微软的SenseCam摄像设备包括一个广角镜头和多个电子感应器，其中有光照强度和光照颜色感应器、被动红外探测器、温度感应器以及多轴加速度仪等。SenseCam可以通过程序设定其拍照的固定时间间隔，或是根据佩戴者或环境变化，通过感应器记录，触发拍照。

如今，谷歌公司正在积极推广他们的生命记录设备——谷歌眼镜。它将平视显示屏、摄像头、麦克风、微处理器以及无线模块等，全都集成在一副眼镜的框架上。通过语音控制，摄像头就可以拍照或摄像。谷歌眼镜的显示屏是联网的，因此可以显示短信或用来进行导航。开发者们还可以为谷歌眼镜开发游戏、提醒软件等各种新的应用。虽然这款产品还处于研发早期，但已经获得了很大的关注。没有人知道它在技术圈之外将会有多流行，但它可能只是这类可穿戴设备的一个先遣部队而已。

其他的小型可穿戴相机还有GoPro和Contour，都是为极限运动而设计的。这两者都有多种类型的基座，以适应不同的使用方式，如固

定在头盔上或是用安全带绑在胸部。另一种叫Loocxie的轻巧型摄像头则可以被戴在耳朵上。

有一些在线博客积累了不少关于这些生命记录设备、系统和软件的信息，包括本章前面提到的那些。戈登·贝尔和吉姆·戈梅尔的“全面回忆”以及凯文·凯利的“量化自我”都是关于生命记录的流行博客。微软公司的戈登·贝尔是一位多产的生命记录者，他已经为一个叫作MylifeBits的项目记录了数年的个人数据。这个项目可以在线查看，图片、视频、电话通话记录、个人信件以及问候卡片等都被分类记录在此并支持查询。

想全面记录你的生活，刚开始最好能有一些引导。有时候人们从关注自己生活的某个方面开始，比如睡觉和起床的时间，然后扩展到其他比较容易记录的方面。哈佛大学的“幸福追踪”项目（Track Your Happiness）是一个关注单一指标的记录系统。参与者会收到系统自动发送的短信，通过短信可以链接到一个简短的问卷调查，询问参与者当前正在做什么以及感受如何。在连续参与几周每天数次的问卷回答之后，参与者会得到一份将参与者的回复形象化了的“个人幸福报告”。

另一个名为“你的数据流”（your.flowing.data）的系统可以让人们通过推特消息，将自己正在做的事情直接发送至一个在线数据库。在发送的推特上附加预先设定的一条“阅读X”或“观看X”的标签，可以帮助精确记录和分类。该系统可以记录消息发送的时间，同时形成时间戳记信息，之后可以在该网站上生成可视化的活动频率图。

“天天日记”（DailyDiary）网站则是向用户推送邮件询问特定的预先选择的问题，如“你今天怎么样？”或是“你今天吃了些什么？”用户通过回答这些问题，获得相应积分，从而可以参与一个在线社区，查看其他人向各自目标努力的进展情况。

尽管有这么多的工具可以让人们通过数字方式记录和组织更多的生活信息，但这些工具尚缺乏易用性，使用起来并不容易，而且很容易让人忘了使用。生命记录的目标是尽可能多地采集信息，只要生命记录中还有很大比重的数据需要依靠手动输入，它就仍将是一个边缘行为，因为记录的目标是获取尽可能多的信息，而大多数人只是选择性地偶尔记录他们的生活事件。

人们有越来越多的方式来追踪和记录自己的生活和习惯，并且可以在后台被动地完成大部分的追踪工作，从而省去了人们输入信息的烦恼。一些人也许会认为，个人数据采集会真正成功，直到人们不再需要不时关心设备状况、数据输入和调整设置；另一些人则相信人的主观性应该始终是数据采集链中重要的一环，这样一来，他们才能自己选择哪些数据应该以什么方式在什么地方被存储下来。

就目前情况来看，个人数据采集技术在一段时间内，将采用一条介于上述两者之间的路径。每个人对自己个人信息的公开程度都有各自的容忍度，所以最好的方式是，追踪服务供应商提供简单易懂的隐私选项，由每个用户自己来进行选择和设置。另一些对隐私不敏感的用户也会感到满意，因为有各种设备来帮助他们自动完成数据采集的工作，他们只需在想要回顾某个事件或想法时再去查询相应的记录就可以。

如今的软件开发者们比以往任何时候都更需要了解人类心理学和社会学。他们需要了解，是什么因素促使人们去记录生活中的一些特别事件或活动，以及如何才能使记录成为人们的日常习惯？他们需要了解，如何将被动式后台记录的数据展示给用户，而不显得唐突和具有侵扰性？他们需要学会，如何让记录生活的设备造福于人，而非平添负担？而对于生命记录设备的开发者来说，他们还需要知道，仅仅是这类技术的存在这件事情本身，就可能极大地改变社会动态性。我们尚不清楚，一旦人们知道自己所言所行都可以被头戴谷歌眼镜的朋

友和陌生人，在不易察觉且未经允许的情况下记录下来，将会做何反应。下一章将讨论本章中涉及的一些敏感问题，并探讨数据在个人层面的一些具体应用。

---

1. 时间戳记（time stamp），是指唯一的标识某一刻时间和日期的一个字符序列。——译者注

第二章
如何充分利用个人数据?

即便一个人的所有信息数据都可以被采集，依然还存在这样的问题：这些数据可以被用来做什么呢？本章将为这一问题提供一些答案，包括几个具体项目的介绍，这些项目的目标都是为了建立一个可以让人们生活得更加健康、更有乐趣的系统。当然，这些项目也都还处于各自的早期阶段，对于个人层面的数据挖掘也还停留在较浅层面。

一类激动人心的应用是，使用个人数据分析来为个人行为提供指导，督促人们改变行为习惯。比如，提醒你在公司的会议上更多或更少地发言，控制你在某个网页上的浏览时间，甚至是督促你戒烟。市场上已有的计步器、热量计等设备已经可以反馈有关健康习惯的信息，试图让人们更加关注所吃的食物，并从行为上有所改变。这种督促被付诸实践的程度，将会决定一个人是否真的可以实现他的目标。

个人数据分析也可以被整合到一些特定的系统中，当判断出有些不太寻常的、可能有潜在危险的事情发生，或是财物有损坏或被偷盗的风险时，这些系统可以发出通知。比如，可以追踪用户的位置以判断安全情况，或汽车是否被盗等。

本章提供的个人层面数据使用项目的列表还远未详尽。更确切地说，本章只是对那些企业家和研究人员已经或将要开始思考他们自己计划的领域做了一次探索。这类项目目前呈现出爆发式增长的态势，每天都有很多关于个人数据的应用程序出现，其中一些具有不错的前景，有些则可能是昙花一现。但是越来越多的应用都指向同一个发展趋势：使用个人数据来为每一个人服务。

暂且不谈这些应用，基于个人数据挖掘的这些项目和产品，都清楚地表明我们需要更加注重隐私保护的问题。哪些人可以获得某个人的数据？数据是属于那些生成数据的个人，还是属于那些数据采集技术的所有者？目前这些都尚无定论。

汽车保险公司是否可以使用人们驾车前往的目的地的数据，从而收取更高的保险费用？已经有一些健康保险公司向穿戴计步器的用户提供保费折扣，因为他们认为这样的用户会进行更多的运动，从而有可能减少身体健康方面的费用支出。健康保险公司会追踪更多其他种类的个人数据吗？比如用户的吸烟习惯或他们的社交网络？信用卡公司也已根据不同用户在某些特定商店购物的消费情况，向其中一些用户收取更高的透支利息。然而，这样是否公平呢？获取这些数据的公司会比生成这些数据的个人获益更多吗？

这些问题都还有待进一步讨论，但在本章中，我们将探讨个人数据采集中的隐私和政策，以及这些方面未来可能的发展趋势。另外，我们还会探究如何让人们尽可能地控制个人数据被他人查看和使用的情况。

可以帮你戒烟的手机应用

当你不经意间养成了一个坏习惯时，对于形成这个习惯的环境，你可能只有一些模糊的或是不完整的概念。但是，如果你能够将这个坏习惯与量化的行为、位置和社会状况等数据关联起来，也许你就有办法改变它。

吸烟行为的研究是一个很好的例子。该项目还处于早期阶段，它希望通过将行为、位置和社会交往与吸烟行为进行关联，试着通过改变人们的生活习惯来改善其健康状况。一旦这些关联因素被获知，工

程师们就可以开发一个手机应用，来识别潜在的诱使人们吸烟的环境。该应用可以适时发出反馈信息，比如提醒你嚼一块口香糖，以代替可能的抽烟行为。而且我们有理由相信，在吸烟研究中开发的方法具有被应用到其他的行为改变项目中的可能性，如一般性药物滥用和依赖、高风险性行为、营养平衡以及日常锻炼等。

由内森·伊格尔和他的同事们开展的这个研究项目的研究对象是18~25岁的年轻烟民。这些吸烟者中，有些人的吸烟频率低达每个月只吸一两根烟。但是近期的研究表明，即使如此低的吸烟频率，长期来看也可能成瘾。而将研究对象设定为年轻烟民，是因为这个年龄段的吸烟者比例高达38%，这一数据超过所有其他年龄段。

据估计，偶尔吸烟者中有一半人在大学四年毕业后依然吸烟，但另外一半则彻底戒烟了。这表明对于大学期间偶尔吸烟的人来说有一个转变期，在此期间，他们有可能在这个关键时期转为不吸烟者，而这正是手机应用可以介入并发挥作用的时期。

使用手机应用来帮助人们戒烟并不是一个新奇想法。事实上，广泛开展的禁烟运动就是通过互联网和移动设备，向青少年和年轻人发送禁烟教育和治疗信息。然而，伊格尔和他的同事们开展的这个名为“生态瞬间评估”（Ecological Momentary Assessment）的项目，其目标是使用手机采集实时数据，并评估同龄人的关系和交往对个人吸烟习惯的影响，以及促使其戒烟的可能。这个项目结合了麻省理工学院一个早先的现实挖掘项目的相关元素，利用手机后台应用程序针对正在发生的个人行为通过简单问卷调查。

就这个项目来说，装有传感器且在年轻人中普及度很高的手机，是一个非常合适的研究工具。该项目计划对纽约城市大学的100名学生进行数据采集，这些人每月至少吸一根烟。手机自动采集并上传各种行为数据，包括通过GPS获得的位置和移动数据、通过通话记录获得的联络数据以及通过蓝牙获得的与他人的空间接近度的数据等。

除了这些自动感应设备获取的数据以外，研究参与者还被要求通过手机回答一些问题，包括吸烟行为、引发他们吸烟的社交和环境氛围、吸烟的迫切度，以及如何应对吸烟需求等。因为在回答这些问题时，手机仍然在同步记录个人的社交网络信息，将这些信息结合起来，可以针对个人及其吸烟的影响要素形成一个生态有效的描述。从而，利用这样的信息分析结果告知人们有哪些有效的社会支持可以帮助其戒烟。

这个项目的终极目标是形成一个预测工具，从而对吸烟者生活中影响其吸烟行为和习惯的因素进行分类，而且其精确度是一般的人工观察无法达到的。为实现这个终极目标，首先需要建立一个可以识别每日、每周乃至每月的个人日常行为结构模式的算法。之前很多研究项目都提出了类似的想法，即用显示个人日常活动及其可预测性的方式对个人行为进行描述。当人们将手机用于建立模式之后发现，该方法比以往的任何方法都更具精确性。

这项研究意义深远，从实践角度来看，它可以在临床研究中被用于回答与社交网络相关的问题，这在以前只通过观察获取数据时是无法回答的。未来临床医生将可以识别出一个人的社交网络中，帮助哪些人最有可能使其戒烟。此外，该研究生成的信息还可以改善手机的干预方式，包括可以通过手机识别用户所在的场景，从而提醒用户，促使其远离可能诱发吸烟的场合。

这个项目在理论研究层面也处于前沿，可以为社会、环境和心理决定因素之间的复杂交互研究提供新的灵感，这些都是研究者们多年来认为可能影响年轻人吸烟习惯的因素。例如，关于年轻吸烟者是否是受到吸烟朋友的影响，还是与他们的朋友们在同一时间段内各自分别养成了吸烟的习惯，一直都有争论。而区分这两种不同的情况，对于设计有效的干预手段来说是非常重要的。

最后，这项通过手机进行的研究中所采用的方法，将可以被广泛应用到其他关于习惯的各项研究中。事实上，研究者们已经在不同的行为问题研究中探索类似的解决方法。在埃森哲公司的一项研究中，工程师们开发了一个应用，可以监测人们在一次会议或是谈话中的讲话时长。这个系统通过用户手机上的麦克风采集谈话数据并存储在一个中央服务器上。该项目的目标是让人们更加了解自己的发言习惯，因为发言往往是难以被自我掌握的。然而，更重要的是，这个系统还整合了手机推送功能，它可以根据实际情况，判断用户发言时间或是沉默时间过长，从而发出提醒。实时分析和反馈机制可以帮助人们更加有效地进行社交和商务活动。

类似的，计算机上的行为监测软件也可以用于改变一个人的坏习惯。第一章中提到几个软件，例如Slife、RescueTime、Klok、SlimTimer，以及WorkTime，可以监测用户在不同应用程序上所花费的时间，并为用户提供反馈报告。这些软件中，有一些还可以提供可视化的分析，显示用户的计算机使用习惯，用户可以据此为自己制定具体目标，如某些特定应用的使用时间和方式。尽管有些软件已经可以通过分析获得数据，用来制定促使你改变使用习惯的策略。但其实人们可以利用这些数据做得更多。例如，一组员工可以在某个工作任务中直接看到其他人的数据并相互竞争，形成一个社交性的博弈，而不仅仅只接受机器的定时提醒。生产效率监测工具是一个最好的例子，作为一个发展成熟的领域，它需要工程师们开发更多的创新工具进一步提升生产效率。

老人走失、汽车被盗前预警

另一类使用个人数据的方式是建立一个监测系统，当某人的行为严重超出常规，并对触及可能的有害或危险领域时进行预警。自然，

这种紧密监视人们行为，并且对反常行为进行预警的想法会引起个人隐私问题关注者的警惕。然而，如果可以在符合伦理要求的情况下使用这些系统，它们可以在一些场合发挥很大作用，比如对偷盗汽车的行为进行预警，或是提醒护士、病人有异常状况。

这一类系统的关键之处在于采集一系列关于个体的位置和社会交往数据，并基于数据分析来判断个体的固有行为或是一般行为的范围，第一章对此已有讨论。

以护理监测老人为例，人们可以通过固有行为来判断，某个事件的发生是否超越了被监测人的行为空间。一个人的行为空间，本质上是其行为的数学表达，包括一系列相互依存的变量。所谓超越行为空间的事件，并不单纯是某人之前没有做过的事情。而是根据相关数据对某个行为进行分析，考虑到先前的行为特征，发现这件事发生的概率太小，以至于极有可能发生意想不到的状况或具有潜在危险。

设想一下，如果一个老人在晚上11点上了一辆公交车，前往城市里他之前几乎从未涉足过的一个地方，并且他的社交圈中也没有认识的人居住或经常会出现在那里。那么，掌握这个老人社交信息和移动数据的手机或其他穿戴设备就可以判断，这个老人的行为是否超越了他的一般行为范围。根据这样一个监测系统的设置，这个老人可能会得到系统推送的通知并被要求进行反馈，或是老人的监护人会收到警报。

包括通用电气以及一些小的初创企业在内的许多公司，都试图在快速增长的老龄人群监测系统市场上占据一定的份额。很多系统使用的都是用在住宅周围安装分布式感应器的方式对一个老年人进行监测。尽管这些感应器可以高精度地感知一些特定的行为（比如正在厨房备餐），但它们的安装和维护较为复杂而且也很昂贵，因此可能会限制该领域和相关公司进一步的发展。

像这种基于固有行为的系统也可以用于汽车防盗。虽然现在市场上已经有多种无线电汽车定位系统，包括LoJack和通用电气的OnStar，它们可以在汽车被盗后对其进行定位，但基于固有行为的系统可以更具前瞻性。当汽车在非正常时间被驾驶到非正常地点时，这类系统会做出判断并向车主进行确认，以确定车主对汽车的这个非固有行为知晓。

大数据，大隐患

当然，所有的现实挖掘程序在理论世界里都是完美的。现实挖掘研究项目能够很顺利地进行的原因之一就是，研究对象们对项目的要求充分理解，他们也相信研究人员会小心严谨地保护他们的隐私。而且他们也知道该项目会结束，他们可以选择自己的数据被采集并被分析，也可以随时选择退出而不会有任何损失。

然而，从科研到现实世界的转换是微妙的。如果是一个第三方公司在获取个人数据并提供反馈，那么如何才能保证隐私得到恰当的保护呢？有意思的是，由于移动电话所具有的基本特点，个人层面数据采集的隐私保护相对来说可能是最容易的。被采集的数据可以通过多种方式存储到云端或企业设置的远程服务器上，但还有其他的选择：手机可以将数据存储在本地。因为手机的处理器能力和存储空间每两年就能翻一番，因此就在前两年，一个小小手机在计算能力上已经相当于体积上比它大很多的个人计算机了，现实挖掘的任务完全可以在移动设备上进行。不断增长的手机计算和存储能力，使得数据不必非得上传到远程服务器上才能进行处理、存储或分享给第三方。这就是说，目前大部分的数据分析，采用的是本地手机和远程服务器共同分析的混合方式。然而，根据1986年通过的《美国电子通信隐私法案》，在某些情况下，执法部门和政府部门不需要获得任何授权，只

需要一张传票，就有权获取远程服务器上的个人数据。但是，如果想要获得存储在个人计算设备上的数据，则依然需要授权许可。

即便如此，绝大多数提供个人数据采集服务的公司，仍然选择本地和远程服务器的混合方法。存储在远程服务器上的数据可以被用于分析以改进产品。另外，这些数据本身对于公司来说也具有很高的价值，可以用于定位目标客户需求。所以，尽管从技术上来说，用户具有了掌握数据的可能性，但实际上他们可能根本没有这个选择权。此外，想要搞清楚数据的存储和处理是在本地设备还是在远程服务器上进行的，需要一定的技术窍门，对于智能手机来说更是如此，因为智能手机的应用总是在后台进行不断的网络连接。而要想搞清楚所有使用程序的服务条款，则需要很强的心理负荷能力：哪些数据是与其他公司共享的？哪些会让数据失效？服务条款是否有更新？更新了哪些内容？如果你对这个程序的服务不甚满意，是否可以将数据转入至其他程序？

如此说来，如果在哪个领域内可以建立坚实的数据所有权的条款，那么必然是在个人层面。因此，数据所有权和隐私保护的新法规需要由企业、用户、团体组织以及法律专家共同来制定。毕竟，我们很容易想象出那些个人数据的使用和分析可以被用来控制、胁迫或勒索用户的情况。当健康保险公司发现一个客户正与可能会增加其吸烟可能性，同时减少其运动可能性的朋友外出，保险公司是否可以据此提高保费？如果软件可以通知虐待老人的护工，他/她所看护的老人正在前往一个更安全的地方，那么到底谁是技术的受益者呢？警察在被控告跟踪别人或实施家庭暴力时，是否有权追查他的行踪？汽车保险公司是否可以根据客户经常开车前往的目的地的安全程度，来收取相应的保险费用呢？

过去几年，在地方政府处理大量个人和公共感知设备相关事务的过程中，所有以上提到的问题都得到了极大的关注。其中有些问题的

答案已经开始有些眉目了。在接下来的几个部分中，我们将简要介绍个人数据层面三个重要的应用程序，从中可以看出，在某些领域，我们迫切需要一种新的数据隐私和所有权形式：比如健康保健奖励机制、现驾现付的汽车保险，以及对家庭暴力和护工虐待行为的防范等。尽管这些只是对个人数据采集可能带来的问题的初步探讨，但作为一个良好的开端，它们可能会帮助我们找到可以广泛应用于不同领域的解决方案。

健康激励是与非

对拥有不健康生活习惯的客户进行经济处罚的做法并不常见，如今最常用的是一种所谓的健康激励项目，即参与者如果选择了更好的生活习惯，就可以获得奖励，如果他有坏习惯不给予经济处罚。事实上，过去的几年中，“健康激励管理行业”出现了越来越多的运营公司。这些公司帮助雇主激励其员工参加康体活动，促使他们的生活方式向着积极健康的方向转变。

这个领域的公司包括Virgin Health–Miles、RedBrick Health、Tangerine Wellness等，它们帮助雇主调解其员工的健康保险理赔，从而最终减少雇主的支出。雇主可以考虑如何利用节省下来的这部分支出来激励员工选择健康生活，一些雇主选择降低员工的自付额，另一部分选择减少保险费或共同付费额度等。

总体来说，有两种项目类型：一类是对参加健康相关课程的人进行奖励，如一个戒烟辅导等；另一类是对达到某些健康指标的人进行奖励，比如在一定时间内减重30磅（约合13.6千克）体重。这两种类型的案例中，参与者都有可能获得每个月100美元及以上的健康保险费用减免。

计步器、手机上的加速度仪、心率监测仪以及数据记录网站等感知端，在这类健康激励项目中扮演着日益重要的角色。例如，有的项目使用计步器来设置员工可以参加的挑战，还有的项目使用这些感知设备帮助人们设置健康目标。如果某个人在某个项目中的表现不佳，也不会被罚款，只是他会因此失去了一个节约保费的机会。是否参加这些项目全凭自愿，但它们确实正变得越来越受欢迎，有些公司员工的参与率甚至高达80%。

使这些项目具有自愿性，并且以“为自己省钱”的概念来代替“罚款”，帮助项目在近些年获得了广泛的参与。在2007年和2008年，包括百得和惠而浦在内的公司都开始对吸烟的员工处以罚金。2008年，惠而浦解雇了一名为了逃避每年500美元的吸烟罚金而撒谎的员工。然而，这个事件至少使得一家公司——Tribune公司放弃了对吸烟员工进行惩罚，这表明有些公司更愿意选择“萝卜”而非“大棒”政策。

尽管这类健康激励项目已经存在很多年了，但未来这类项目的数量以及其中使用的技术的复杂程度都将提升很多。感应器不仅越来越便宜，体积也越来越小；更重要的是，立法也在推动这个产业的发展。2010年的春天，美国总统奥巴马签署了一个推动该行业发展的法案，包括增加提供健康管理项目的企业数量、提升员工参与人数、更加有效地追踪项目以及提升项目的整体效果等内容。毫无疑问，实施健康激励项目的公司从中获得了财务上的回报。美国健康理事会称，在健康激励项目上多投资1美元，在医疗保健的支出上就能相应节约3美元。

但是，如何保护员工的健康隐私呢？根据2008年《美国残疾人法案》（ADa）修正案，雇主不得向员工询问其医疗健康状况，除非是工作相关并且有业务上的需要。然而，美国平等就业机会委员会（EEOC）2002年的一份实施指南中称，即使不与工作相关或符合业务需要，雇主仍然可以了解员工的医疗历史，作为健康管理项目的一

部分。1996年的《健康保险携带和责任法案》（HIPAA）及2006年的后续法规，允许雇主在员工健康计划中安排多样化的保费、免赔额以及共同负担保费的支付标准。《健康保险携带和责任法案》中的反歧视条款，禁止对健康状况相似的员工收取不同保费。但现在还不清楚《健康保险携带和责任法案》及一些新规中的类似条款，如何与美国平等就业机会委员会要求的自愿原则相协调，并与《美国残疾人法案》中的隐私规定相一致。

这类项目的另一个挑战是如何保证基本的公平性。早期相关法案的一个重要规定是奖励额度不得超过员工保险额度的20%，但2010年奥巴马签署的新法案将比例限额提高到了30%。有些机构认为这一比例过高，可能会对低收入者造成过多负担，因为低收入者相比高收入者更可能出现健康状况。

当越来越多的雇主开始实施健康激励项目以应对逐年上涨的健康保险费用时，这些问题将逐渐浮出水面，并成为中心议题。并且，当人们能够通过更加精确的感知设备和算法，来判断某些特定行为与健康状况好坏之间的因果关系时，这些问题会变得更加复杂。

车载感应功与过

如今，很多新车上都装配了GPS和其他导航设备、用于辅助泊车和倒车的外部摄像头、探测车辆间距离的感应器，以及远程控制锁车和点火的控制器。在某些情况下，这些系统能使警察快速地追踪到一辆被盗汽车，甚至远程控制将车熄火。

随着汽车的控制系统愈加计算机化，我们可以合理地推测，未来的某些车型中可能会安装可以提供驾驶习惯的一般性反馈的感知设备。实际上，现在想监控驾驶过程的话，人们只需要购买并安装一套

基于GPS的驾驶监测追踪系统就可以了。有些保险公司招募志愿者，并使用这类车载感知设备来评估他们的安全驾驶习惯，据此对其保险费用的收取标准进行调整。

传统的汽车保险根据驾驶者的过往驾驶经历计算保费。衡量一个驾驶者是否“安全”包括很多方面的考量：有记录的交通违规、年龄、所驾驶汽车的类型等。然而，近年来发展出了一种新的模式，被称为PAYD（pay–as–you–drive,“现驾现付”）。

PAYD系统不需要一个驾驶者的过往驾驶记录，它使用的是驾驶者驾驶习惯的持续反馈。简而言之，PAYD根据汽车的里程数据，对驾驶频率较低的驾驶者降低其保费的收取。但是车载感知设备可以获取更多更精确的数据并反馈给保险公司，包括速度、距离、每天的开车时间，有时则是一些GPS数据。该系统可以让保险公司针对参保人超速行驶驾驶时间过长，以及经常开到犯罪多发区等情况征收较高的保险费用。

美国的很多保险公司，包括Progressive、Liberty Mutual、MileMeter，以及GMAC，都采用PAYD方案，少数几个英国、加拿大、南非和日本的公司也提供一些PAYD选择。在美国有超过30个州认可PAYD方案，但由于客户的需求不多，该方案在美国以外的国家被接受的进程一直十分缓慢。

例如，加利福尼亚州在2009年就推动了PAYD保险方案的立法。除了可以为一些司机降低保费，PAYD方案还提供奖励政策以减少人们不必要的驾车出行，从而降低温室气体的排放。据估算，如果PAYD保险方案能在全美范围内达到30%的使用率，将可以减少10%的汽车行驶总里程，这意味着可以在10年内减少大约5 500万吨的二氧化碳排放量。

在加利福尼亚州推出的最初方案中，保险公司可以要求用户在他们的汽车上安装电子监控设备。因为里程数据调整起来很容易，所以这种方法与仅依赖汽车里程表相比，可以减少用户欺诈的可能性。但是，一个非营利性的数字版权倡议法制组织——电子前沿基金会迫使立法者对该法案进行了修改，将里程表和电子监控仪的数据都纳入可用范围。

对PAYD系统的担心，主要围绕保险公司可能会将用户数据出售或共享给其他机构或政府，还有就是这些系统的安全性。此外，消费者在怀疑收费不合理时可能难以对保险公司账目进行审核。例如，电子监控系统可能依赖多个变量，包括速度、位置、驾驶时间和距离以及驾车频率等，来衡量和计算保险费用。但是每个变量参与计算的具体权重，可能由于知识产权保护的某些规定使用户无法对自己保费的完整性进行审查。最后，系统可能并无法全面反映驾驶者的驾车习惯。一个偶尔超速的驾驶者，实际上，可能比一个在车流中总是减速和犹豫的驾驶者更可靠。

所以，尽管一套智能感知设备和算法可以帮助人们更快找到丢失的汽车，也可能帮助降低某些驾驶者的保费，但它无疑也使隐私政策和法律的制定变得更加复杂，最终使一些驾驶人处于不利的位置。

无法回避的个人隐私

每年都有更多的人自愿将其个人数据在线分享，有的是锻炼路线，有的是不经意的想法，有的是自家孩子的照片。尽管那些在线分享个人数据的人们，也许并没有过多考虑他们数据的潜在使用方式；他们只是分享数据，并且在大多数情况下，他们可以在自己希望的时候，从公众的视野里将这些数据移除。

但是，对于一位老者或是一个精神或身体上不健全的人来说，情况就大不一样了。这些人通常对应用在他们身上监测其位置和行为的数据感知系统没什么话语权。其结果就是，他们只能选择相信那些看护他们的人和系统。

有些人认为对老年人或身体功能受限者进行监控的想法存在一个基本的道德问题，即该做法威胁了他们作为人的基本尊严。当看护者或医生认为一个人需要被监护和追踪时，病人的自主性就被削弱了，其基本公民自由的权利也更容易被忽视。

针对这些基本的人身自由权利，看护者在照看病人时逾越道德界限的情况并非鲜有听闻。不幸的是，并非所有看护人员都能很好地照顾病人。在美国一些地方，看护者虐待病人并进行经济剥削的案例越来越普遍。这促使联邦立法者们制定相关法案，要求对看护者执行更为严格的背景检查等措施，以减少看护者虐待老年人的事件。然而，由于大约90%的施虐看护人都是病人的家庭成员，这种严格的背景检查措施是否会有效果还很难讲，因为施虐者可能完全没有作为看护人的历史记录。

与此同时，技术的进步也使得老年人和身体功能受限者可以独立生活得更久，看护人员只需要远程看护。例如，通用电气公司开发了一套产品，利用无处不在的感应设备可以精确但不被察觉地识别独居老人的行为。当前，简单的行为追踪系统更为普遍，比如基于GPS的“地理围栏”可以在被追踪人离开预定空间范围时向看护者发送警报。这款产品是为阿尔茨海默症患者所设计的，但也可以用于儿童和青少年的看护。

立法者应该尤其注意那些可能助纣为虐的设备，比如i–TAG和其他一些更复杂的感知系统。它们会在一个被施虐者逃离时，立即向看护者发送提示消息。并且如果通话记录也被设备或软件监控着的话，被施虐者就更难逃脱了。

在某种程度上，与施虐看护情形相似的还有更加普遍的家庭暴力，甚至是陌生电子追踪。在汽车上偷偷安装追踪设备已经是较为简单的了。至少曾经在一个案件中，家庭暴力的受害者起诉了Foxtrax汽车追踪设备制造商，称其设备在一次攻击行为中协助了施虐者。

警方也在防范家庭暴力中使用个人GPS追踪设备以彼之道，还彼之身：美国有越来越多的州开始使用GPS设备监控身负限制令的嫌疑人，这种方式之前被执法机关用于监控被指控实施儿童性虐待的人。

尽管我们都知道使用个人数据的设备可以帮助个人获得更高效的生活、加强锻炼或是戒除烟瘾，这些设备有很多令人激动的用武之处，但是隐私却是它们无法回避的问题。研究人员试图设想这些设备运行多年后的情景，以便更加深入地讨论使用过程中必然会出现的问题，还有随着这类技术进步不可避免会产生的影响。然而，一项新技术中关于个人隐私或技术本身的局限性，比如基于GPS的手机应用或是基于RFID的收费芯片，往往只有被大量接纳使用之后才会显现并开始引发问题。

针对人们对隐私问题的担忧，一个最普通的方式是让人们可以自由选择进入或退出某个项目，使用或停用某项技术。然而有些时候想完全退出行不通。因此，需要使用数据采集和分析方法从而为人们提供多种隐私保护选项。每个需要用户进行设置的隐私选项都应该被准确解释，且清晰易懂。例如，很多人会在程序询问时允许手机应用使用其位置信息，但是很少有人会清楚知道这些位置数据是如何被使用和分享的。

人们对隐私的态度也在不断变化，在脸谱网上安心发布自己照片或者在推特上发布感想的人们都会同意这点，公共生活和个人生活之间的界限已经日益模糊。那些发布到网上的图片和文字可以被任何人浏览和搜索到。那些即将出现的使用个人数据的新技术，无论它们的

数据是否公开，都极有可能利用公众认同或至少是公众度等以减少为不同应用采集和处理个人数据时产生的摩擦。

这些个人层面的现实挖掘方式激动人心，它们用数据来改变人们的习惯、提升其健康状况并节省支出。随着有关个人“大数据”的想法日益普及且对大众来说日益重要，政策制定者、社会活动家和技术专家们关于数据所有权和某些基于数据的产品的合法性讨论将会越来越多。现在，是时候来认真对待基于数据的产品可能引发的问题了，正如我们之前提到的那些。如果能以正确的方式来使用这些产品，那么它们在未来将会为后续产品提供积极而有益的范例。

REALITY
MINING

Using Big Data
to Engineer a Better World

第二部分
数据驱动下的社区和组织

挖掘社区中人们的行为，并把它们和其他诸如交通、空气污染或其他环境指标数据关联起来的工具，对于关心社区的积极分子来说会很有用。为了改善社区，积极分子可以向在公共空间活动的人们提供空气质量感应器，比如那些在公园游玩或在等公交车的人。如果这些感应器检测到空气污染物水平较高，那么就可以发起针对当地工厂的请愿运动，要求他们减少污染物的排放。

第三章
群体的数据获取

正如第一章中提到的，记录单独的个人信息的低成本商业方法有很多。它不仅方式相当简单，而且被记录数据的人们在某种程度上可以控制自己数据的使用方式，以及怎样使这些数据为自己带来收益。但当人们需要从小规模的群体中采集个人信息时，采集数据提供适当奖励以获取更多数据的复杂性会提高，即使这些人效忠于同一个组织或对象，或者有着相同的目标。

用来进行个人分析的数据，其采集通常给人的感觉较为私密和保守，因此即使在较小范围内分享所采集的个人信息，也会让人们觉得隐私被暴露或者缺乏安全感。从另一个层面来说，小群体的数据采集

与大群体（一般超过1万人）相比有较大差别，因为在小群体中采集数据，其个人的身份信息与其数据往往是明确联系在一起的，而针对大群体的数据采集，个人的身份信息一般会从数据上被剥离，以给人匿名的感觉。

当涉及具体的数据使用案例，如公司或社区组织使用个人数据，一些特别的挑战便会产生。当某人的上司或其他权威人士可以获得具体的个人信息，人们会本能地排斥。同样，社区层面的数据获取也很困难：城市规划师、政客或社区管理员等需要使用社区层面个人数据的人，可能无法获得社区中每个人的支持，从而导致只有部分人自愿参与。

因此，大部分该层面上的数据采集，往往局限于参与者自愿签署了数据共享协议的研究项目。因此，这类项目一般具有一定程度的同质性，并且通常是在大学校园里开展，因为这里的人一般具有共同的行为特征。例如第一章中提到的，在2003年开展的“现实挖掘”项目，其中的100名参与者就全部都是麻省理工学院的学生。

尽管商业领域已经为采集小群体规模的行为数据做了一些尝试，但收效缓慢。有一个例子是“智能”会议标识卡，这些智能卡基于与会人员的社交模式通过RFID技术或红外线感应器来分辨他们的社交网络。这些智能卡为人们提供了共享交流信息的一种简单方法，也使会议组织方能够很容易地得知每个分会场的参会情况。尽管这些智能标识卡已经出现了很多年，但仍然未被广泛使用，一部分原因是其花费较高，还有部分原因是其可靠性及应用范围有限。

在工作场所，人们并不缺少可以用于存储和挖掘工作人员数据的软件，包括生产的内容、网页浏览以及邮件发送。这些作为“知识管理系统”为我们所熟知的软件工具，承诺通过允许员工更便捷地获得其组织内部信息或知识的办法，来提高员工的工作效率、促进团队协作以及交流。然而知识管理工具的成功很有限，因为将它们融入一个公司

的技术体系及其团队文化是很困难的。“知识”通常是被另外附加到一个系统上去的，但只有经过这种组合，它才会真正对某一个特别的项目产生用处。另外，知识管理工具的维护费用较高，人们也无法及时对其进行衡量。

另一种监控工作场所的类型是通过对刷卡钥匙监视器，与工作相关的手机通话（词汇、语调、频率等），以及监控办公室或设施附近人员移动的感应器进行分析。但是这些工具几乎从未被公司采用过，唯一的例外就是呼叫中心，由于显而易见的原因，其通话往往是被监听的。很少有公司愿意冒着疏远员工，甚至可能引发法律诉讼的风险，来尝试通过这些有待证实的方法提高工作效率。

在公司和工作场所之外，研究人员发现一些在社区层面搜集数据的有趣原因。一些在大学校园及其周边开展的项目，依靠市民参与来探索不同的环境层面问题，包括空气污染、垃圾堆放、道路状况以及通勤路径等。这些数据采集是简单直接的，往往采用手机作为传感器，但其挑战在于，学术环境之外如何拓展项目研究范围。超出学术范围的项目扩展，其中的一个关键是提供一个最合适的激励办法，让人们能够合理地权衡是否要在小范围内共享自身数据以从中获利，并做出有依据的决定。

也许最显而易见的方法是向人们提供智能手机的同时提供流量及通话套餐，以换取他们的数据。几年前，IMMI公司（综合媒体测量公司）就成功地通过提供智能手机的方式来换取用户向他们提供日常生活中的声音片段。该公司的目标是通过这些声音样本来判断用户正在消费的媒体类型，其中包括广播、电视、电影等。这一信息可以用来分析特定媒体内容的受欢迎程度，类似于内尔森的媒体调查系统。值得注意的是，这一技术并不记录和存储个人对话或原始数据。尽管如此，这样的一个项目仍然很难通过大学伦理委员会和机构审查委员会

的批准，因为如果用户没有签署参与研究同意书，那么他们的语音数据就不能被采集，而语音数据采集却是这个项目主题中不可避免的。

本章探索了当研究人员和研究公司试图从特定群体中采集非匿名数据时，所面临的技术、法律以及社会挑战。为什么有些方法成功了有些却没有？以及在这个尺度上的“现实挖掘”还面临着怎样的机遇？

智能标识卡

会议和公开活动是采集不同个人数据的绝佳场合。从参与人注册开始，他们就已经提供了一定量的人口统计学类型的个人基本信息，当他们出席这些会议和活动的时候，往往希望尽可能多并尽可能高质量地与其他参与者进行交流。一个可以追踪参与者的互动情况，并且至少可以提供一份电子联系人名单的智能会议标识卡，相比传统的握手以及交换名片来说是一个实用的改进。

不同的智能会议标识卡在设计、功能以及其采集存储数据的方式上都会有所不同。但最基本的，它们都配有RFID芯片。这种芯片包含一部分识别信息并且可以被特定的识读器在近距离或几米范围内识别。RFID会议标识卡最常用于记录会议的参与情况以及参会人员的用餐情况。但有时也会用来记录与会者的移动路径，还有他们与安装在不同厂商的展台上的识读器之间的距离。这些标识卡将识别信息传输到装在特定位置的RFID电子识读器上，这些识读器按照获取信息的顺序连接远程服务器上的应用程序，用以分析这些数据。

这些RFID标识卡往往和传统标识卡外观无异，轻重相差无几，但其价格仍然是推广的障碍。尽管制造RFID芯片的花费在过去10年间下降了不少，智能标识卡仍然比传统标识卡昂贵。像微软和IBM（国际商用机器公司）这样的公司曾在一些会议中试验性地使用RFID标识

卡，而另外一些公司，如安联科技（Alliance Tech）和会议策略（Convention Strategy），也曾为其做过商业化推广。这些商业标识卡可以记录人们观看某个展览或在某个展台逗留的时间，还可以在一些拥有某个特殊头衔或在特定公司工作的重要人士出现在识读器范围内时，向展台处的工作人员发送电子邮件和文字短信。

这些智能标识卡的一个主要问题是使人们缺乏对隐私的控制。一旦佩戴上这个智能标识卡，佩戴者的个人数据就会一直暴露于RFID识读器之下，除非他用铝箔之类的导电材料物理阻隔信号。人们无法选择一台RFID识读器能够识别什么，也无法决定哪一台识读器能够识别标识卡的事实，造成了这样一种情况——佩戴者缺乏对卡片的控制并可能会因此不愿意佩戴这些标识卡。这也是推广此类技术的一大障碍。

电子标识卡是另外一类智能标识卡，它采用包括RFID在内的多种传感器来让人们分享联系方式，并且可以与他们自己以及其他人的标识卡互动。比如，这些标识卡可以带有一个简单界面，通过显示屏和一些按钮来让人们在设备上查看数据，回答问卷调查。

这种互动性相对于简单的RFID标识卡有一个巨大的优势，因为人们可以选择他想传输的信息类型，这有利于提高个人对信息的掌控感，从而提高其分享数据的心理舒适度。在麻省理工学院的“现实挖掘”项目中，参与者在不希望手机采集他们的行为数据的时候，可以选择“隐身”。隐身模式这项选择可以降低参与者对于针对其个人的数据采集的担忧，校内校外与他们交往的人也位列担忧之中。报告称，参与者对于自身具备掌控数据的权利而感到满意，但在项目进行过程中很少有参与者真正激活了隐身模式。

由于其（装置）相对复杂，装载了传感器的电子标识卡比RFID标识卡更贵，从而阻碍了前者的广泛应用。另外，电子标识卡更重，这使得它们不适合全天佩戴，也不适合日程比较长的会议。

在这个领域中，一家著名公司是麻省理工学院的衍生企业nTag。2009年3月安联科技收购nTag，并雇用了其核心员工，现在安联科技是nTag技术的独家供应商。除了众多其他功能之外，nTag标识卡还可以促进佩戴者与其他人交谈，交换电子名片，与工作人员、演讲者以及其他参加者更容易地交流。与RFID标识卡相似，nTag标识卡同样可以让会议组织者监控会议与展台的参与情况。

现如今，有越来越多的人都在使用具备很多标识卡所提供的功能的智能手机。我们现在尚不清楚，这些标识卡公司将怎样在这样的时代中经营下去。雇用一个软件开发人员来为主流的智能手机操作系统开发应用程序，以便与会人员使用自己的个人设备来实现nTag标识卡所具备的功能，这可能是很划算的选择。

科学地盯紧你的员工

凭借数据挖掘软件而闻名的公司，例如甲骨文和IBM等，多年来一直在运用相关技术对其员工所产生的万亿级别数据进行分析。这类被称为“知识管理”的数据挖掘项目，其基本目标包括规范交易流程、开展高效的营销活动、识别欺诈行为，以及捕捉律师事务所立案所需的“电子证据”（eDiscovery）。最终目标是通过提高员工工作效率，或者更为理想的——提高员工满意度来改善工作环境。

Tacit软件公司是这一领域中的翘楚，该公司成立于1997年，并在2008年被甲骨文收购。Tacit公司开发的软件（第四章中将会详细论述）可以分析电子邮件交流、博客和维基百科词条，来确定人们的兴趣和专长。这一软件用在大公司中可以帮助员工相互认识，从而在解决棘手问题时可以相互帮助。洛克希德·马丁、诺斯洛普·格鲁门，以及葛兰素史克等公司都采用这一软件。

另外，Declara公司目前作为一个所谓的虚拟咨询协作平台，其软件可以创建一个同机构人员的社交网络，并且通过他们所产生的数据为搜索者匹配最佳信息。

Tacit和Declara公司的技术具有高度的自主知识产权，因此可能价格较高，但还有其他较为便宜的开源选择。比如OpenKM兼容市面上主流操作系统，并且文件可以存储于诸如甲骨文和MySQL一类的本地文件系统或数据库管理系统。此外，还有大量处理知识管理各个方面工作的开源软件工具，举几个例子来说，比如赛门铁克公司的KAON框架、Exteca、Haystack和Kowari。

员工创造的数据除了显示他们的专业知识和兴趣之外，还能反映同一机构中工作人员的效率。在过去的几年间，许多公司的人力资源部门都已经接受了使用员工创造的数据来评估员工工作效率的方法。虽然这仍然是一种新尝试，但它使企业管理者能够量化评估一个员工的工作能力以及他在公司的员工网络中所扮演的角色和产生的效力。

Cataphora公司现在借助他们之前开发的，用来依法证实一些公司过去曾经犯过欺诈罪的技术，来帮助现在的公司判断他们的员工本身是否涉嫌犯罪。该软件监控员工的电子事务，包括电子邮件、发送和接收的附件、即时消息以及日程安排。该软件负责找出常规交流模式，并根据该模式找出例外情况，比如签字时使用与平时不同的语言，有时候就是作弊的一种信号。这个软件可以发掘员工之间的小圈子，给人力资源专家提供潜在的线索，使其能够分辨哪些人工作产出能力强，哪些人工作产出能力弱。除了提供企业级软件之外，Cataphora公司也为员工提供一份说明书和一个商业版本软件，这样员工就可以了解他们的老板会怎样破译他的个人数据。

尽管对于一个在办公室工作人员来说，发送和接收电子交流信息是其一天工作中的重要组成部分，其工作效率和社交效果也可以通过

其他指标来观察。例如电子ID卡的记录可以显示一个人一整天的移动轨迹。

另外，文字短信、语音通话，还有通话记录也可以用于评估员工的社交网络及其行为。众所周知，客户服务中心会通过监听语音通话来尝试优化雇员与客户之间的互动。现在的软件可以分析通话内容，让公司了解顾客通话时的情绪。Verint系统有限公司集成了语音分析、电子邮件文本、即时信息以及调查的反馈，以期更全面地了解其用户对公司的满意度。

除了使用客户服务中心的行业以外，雇主监听员工的语音通话的情况其实并不常见。何况，想要了解一个公司员工的社交状况，并不需要完整地记录他们的语音通话；对于鉴别员工个人或团体的工作能力来说，也没有必要这么做。分析语音特性，比如从上一轮谈话转换到下一轮的风格以及面对面谈话的速度，可以用于确定社交网络的结构。将这类研究拓展到语音通话上，可以免除对整个通话进行录音的必要，这不仅能最大限度地减少隐私方面的隐忧，还有一些额外的好处。

在这个领域中对于合法性的担忧是合理的，特别是现行的法律并不完善，缺乏切实保护雇主或被雇佣者的条款。尽管对所有与商业相关的通话进行录音都是合法的，但在大多数时候这种监控需要事先告之通话双方。个人通话，在被认定为私人通话的情况下，是不可以被监控的。然而，如果公司没有明确规定工作时不可以进行个人通话，那么雇员的个人通话就有被监控的风险。

2010年6月，美国最高法院在安大略市诉杰夫·库昂一案（美国最高法院判例汇编索引号：130 S.Ct. 2619, 560 U.S.）中判定，在工作相关的背景下对警局分配给警员呼机上的信息实施搜查是合法的。在这个案件中，城市管理部门试图通过对日常使用情况的调查来决定是否修改与寻呼机提供商签订的无线合约，并因此获取了一份杰夫·库昂的

里面含有大量淫秽信息的寻呼机信息副本。这次判决的依据是美国宪法第四修正案中关于政府搜查公务员的电子通信记录的规定。尽管这个案件和此次判决的结果可能对公务员的隐私空间界定产生了一定的影响，但当雇主是私营业主时，此类案件在有关隐私权的问题上仍然没有定论。

对于雇员来说，雇主对他们的某些监视如果和一定的利益或者加薪挂钩，可能会更好接受一些。对于这样的系统，一个重要的要素是透明：采集和分析数据的方法要解释清楚，得出的结果要给予员工反对和申诉的权利。比如第二章中提到的健康激励系统，一个合适的员工监控系统的输出结果应该只用于激励员工而不是用于处罚他们。

然而，知易行难，职位的上升和工资上调等方面的不一致经常是公司中争论的焦点，远比每个员工卫生保健费用的差别更受关注。另外，研究这些员工数据远比健康激励系统所用的直截了当的方法要复杂得多。例如，健康激励系统可以根据一个人每个月所走的步数，相应地降低一定量的健康保险费用。一个最好的提高员工工作效率的方法是提高他们对工作空间的满意度，但如果雇主侵犯了员工的权利和隐私，则所有的努力都有可能适得其反。

建设更美好社区

当从社区层面考虑采集数据，则会遇上完全不同的挑战和机遇。这些项目一般是基于具有一定同质性的群体来开展的，比如大学生和教授，这样的数据采集项目，往往在地理范围上被限制在大学的附近。然而，这些项目提出了可以提升参与度的潜在应用和激励方案。在一些项目中，研究人员以及一些公司利用人们想要创造出一个更宜居的社区的愿望，来克服参与者不情愿分享个人信息的情况，这个办法在一些情况中还是很有效的。

英特尔和加利福尼亚大学的研究实验室探索了一种新的方法，根据便携的个人空气质量传感器追踪到的数据来绘制污染状况分布图。这一项目的名称叫作“常识”，通过个人和集聚的人群的曝光来采集数据。以迎合人们渴望改善自身健康状况以及周边环境愿望的这样一种方式，研究人员说服了参与者通过自己的手机分享位置信息，以及分享通过穿戴式传感设备追踪的一氧化碳、氮氧化合物、臭氧层以及光线、温度、湿度和空间方位等一系列信息。这种采集的结果可以帮助社会活动家动员社会群体有的放矢地行动起来。

但是，要将这样的应用推广到更多的人群可能还需要几年时间。主要的挑战是，将大多数环境传感器集成在手机上这样的做法缺乏大规模市场需求，因此很难畅销。另外还有个挑战是，由于采集数据的手机不同，如何保持数据的“清洁”也是一个问题。因此，在这类数据采集上的商业尝试，最可能需要的是，使用不贵、不引人注目、带有蓝牙功能且可以和用户移动设备连接的外部传感器。

由德博拉·埃斯特林主持的一些在UCLA（加州大学洛杉矶分校）开展的研究项目，探索了这类只通过手机在社区层面采集数据的方法。恰好，这些装有摄像头和GPS的手机非常有可能为改善社区做得更多。实际上，把UCLA的一些项目参数进行整合的想法，在众多创业公司都在开发类似的手机应用的当下，正好迎合了大众的需要。

其中有一个叫作“个人环境影响报告”的项目，利用安全地定期上传到远程服务器上的数据向人们提供环境影响和环境暴露水平的信息[注]。通过分析用户的位置轨迹，还有加利福尼亚州数以千计的社区的天气状况、估算交通模式等数据，系统可以生成一份报告，估算参与者暴露在什么等级的环境烟雾之中，以及他经过了多少家快餐馆。它还可以通过参与者所开的车，来评估一个人的碳影响以及他对学校和医院这类敏感地点的影响。

另外一个UCLA的项目让参与者通过手机报告大学校园里未被归置的垃圾的情况，该项目说明了人们在社区层面分享数据的潜在好处。在这个项目中，大学生被鼓励将校园中垃圾桶里的内容用手机拍照，以便确定回收垃圾桶应该如何布置才能获得最高的使用效率。

手机在UCLA还被用来寻找更好的自行车通勤路径。一个名为Biketastic的项目，通过GPS轨迹采集骑车人路径，以及通过加速器的数据来记录路面粗糙程度并使用声音数据来帮助判断交通状况或道路施工情况。当和空气质量、时间敏感的交通数据以及交通事故等其他数据集结合起来时，人们就可以全面、清晰地了解一条自行车通勤路径的情况了。

声景监视

正如许多UCLA的项目所证实的那样，手机上大量的传感器使其非常适合作为环境监测终端。但一份简单、连续地采集所获得的音频文件同样可以很有用。2010年，Arbitron股份有限公司取得了IMMI的技术文件、专利权以及IMMI的商标名称并将其改名为Audience Measurement Technologies。IMMI使用周围环境的音频来鉴别人们正在使用的媒体类型。IMMI直接与尼尔森媒体研究公司竞争，试图了解所有形式的媒体，从电影、收音机、游戏到MP3（一种音乐文件播放器）、电视，以及网络。这一技术通过手机或个人电脑上的录音工具来实现。该录音工具在一分钟内数次记录声音片段，并与声音数据库中的信号进行对比。该技术不需要用户的主观参与，因此可以提供一个无偏见的个人声景样本。它可以在用户注册使用该服务时，交叉引用个人信息来提供个人听力习惯的实时报告。

已经有超过一万人同意使用IMMI的技术。尽管这一群体规模使得IMMI的技术超出了本章小规模群体的范畴，但其客户端采集数据的基

本方法，以及它所采集的数据并不是关联到匿名用户或是某个群体的平均值，而是关联到一个可识别的个人这个事实，使得它很适合用于在相对小规模的群体中采集数据。

当一个人决定使用IMMI的技术时，他需要将其下载至自己的手机或个人电脑上，或接受IMMI提供的预装该数据采集技术的智能手机。每隔几分钟，他的设备会采集一份音频样本，紧接着音频会被转换为数字信号，之后再传输至该公司的服务器。IMMI公司的服务器根据监控的各类媒体，如商业广告、预告片、电影和音乐，计算形成各种媒体的目标内容文件，如果手机传来的数字信号与IMMI公司服务器中的信号相符，用户所使用的媒体就可以被识别出来。信号匹配的过程只需要几秒钟。

在个人隐私方面，该公司宣称，语音通话以及其他非媒体的声音会被过滤掉。因为音频信号会立刻转换为数字信号，而这些数字信号是已经与预先识别过的媒体信号进行匹配过的，因此该公司没有理由也不具备分析其他音频的方法，这些不被识别的音频包括对话以及街道的声音等。而且在这一技术的具体应用中还有一个关键点：IMMI向接受预装软件的手机用户提供设备和服务补贴。这至少对于一部分人来说，足够消除他们对于该技术可能监听自己的生活的顾虑了。

IMMI的技术现在属于Arbitron公司，后者研发了便携式人员收视测量仪，该设备只有一张纸的大小，用于监测电视广播信号中人耳不可闻的音调。当这些信号被监测到时，它们会被无线传输到Arbitron的服务器。这种便携式人员收视测量仪现在已经在加拿大、冰岛、挪威、瑞典、丹麦、比利时和哈萨克斯坦等国家得到应用，而其音频编码技术则在新加坡获得了使用许可。

对小规模群体的数据采集来说，最大挑战在于寻找合适的激励方式鼓励人们参与，特别是在个人能够轻易地直接获得自己所产生的数据的情况下。智能会议标识卡的发展可能受制于经济成本，但工作场

所和社区具有使用“现实挖掘”技术的可能性，通过这些技术可以更好地把握人们的实际工作和日常表现。而且，如果在社区中分享数据的好处可以被高效地交流和展示，很有可能会有更多的人愿意接受某些特定形式的数据共享。

IMMI偶然发现了让人们分享他们所处环境的私人音频的方法：他们提供免费的智能手机，并承诺私人对话会被排除在数据采集外。但实际上，人们可能需要对这种方式的成本效益进行更深入的研究。特别是，曾有一个研究发现，学生们为了获得平均约10英镑的补贴而愿意分享自己的位置信息（通过他们的手机）。经常离开学校的学生认为他们的隐私更有价值，相比那些只在学校附近活动的学生，他们索要的补贴更高。这一研究表明，人们确实会为自己的位置数据估价，而且这一价值的高低取决于一些特定的因素，比如移动性。

研究人员和商人需要更加重视个人数据的价值并做出相应的补偿。考虑到人们固有的隐私顾虑，开展研究的群体规模是个微妙的因素，但合适的规模也可以让大量应用获取负责任的参与者。

---

1. 环境暴露水平是指人群接触某个环境因素的浓度或剂量。——译者注

第四章
数据价值如何落地?

正如第二章中提到的，目前已有很多方法可以利用个人数据来开发系统和工具，以帮助个人实现自己的目标，或过上更健康的生活。这一章我们将介绍小规模群体的数据是如何被用于提升个人的工作效率、改善其健康状况，促进更多有效的群体互动，以及建设更健康更宜居的社区的。利用小规模群体数据，工程师可以更简单地识别有意义的或偶然的人际关系、事件以及人们所处的环境，这些可以帮助公司管理者、积极的市民、健康产品提供商以及当地政府官员更好地决策。

在这个规模上，社交网络中的联盟和层级划分现象就变得十分清晰，而且有助于人们理解群体行为。但这些观察只有在涉及具体行动时其关联性才最强，才有用，比如让一个机构运行得更平稳，或者让一个社区变得更宜居。因此，不把它限制在观察科学的范畴中这一点很重要，设计更好的机构和社区从根本上来说还是要落实到实践。当然，设计一个能够合理激励人们分享数据的系统会是一个挑战，它需要能够让人们放心地分享自己产生的数据，而且在特定的人群能够获取这些数据时，人们不会感受到威胁。

让人高兴的是，很多公司和研究人员都在开发这样的应用。例如第三章中提到的Tacit公司开发的知识中介系统，这种系统分析员工的工作和沟通行为，帮助他们与同事建立新的联系。通过一个叫作Serendipity的手机应用，它用相似的原则可以在正式会议或私人约会的情境中帮助人们相互结识。但关键是要保障每个人的隐私，这些系

统只能使人们更容易结识新人，而不是以个人的隐私或安全为代价，强迫人们相互结识。

从工作场所的数据转向社区层面的数据，可以从相对小规模的关心改善社区层面生活质量的当地居民入手，利用他们提供的数据。在第三章中我们也介绍过，UCLA开展了一系列关于参与式感知的项目，让市民能够监控空气污染水平，比如一个可以标示出哪些道路对于哮喘患者来说更为安全的应用。

但更有意思的是那些已经从实验室走出来，投入商业应用，使用的规模扩张到了数千人的参与感知方式的项目。例如纽约大学衍生公司的Noah项目，让市民科学家为他们所观察到的生物体拍摄带有地理标签和时间戳记的照片并对其做笔记，经过一段时间的积累，就可以为当地生态系统的健康状况提供简况信息。另外一个手机应用CitySourced，则是让人们将拍摄到的城市衰败区域的照片发送到各自的市政厅，这些照片同样带有地理标签和时间戳记。CitySourced是一系列可用，且已经被整合到城市工作流程系统和城市311[注]热线（非紧急事件）中的应用之一，人们可以直接把自己采集到的数据分享给能够解决该问题的政府工作人员。

对于社区来说，如果没有被该系统成功激励而愿意提供潜在敏感信息的这部分居民主动提供潜在的有价值信息，这些参与式感知系统就没有价值。随着这类数据分享激励机制被日益优化，这样的系统会在塑造和维持我们工作、生活的场所的行动中扮演着越来越重要的角色。

本章关注的是这些本地应用程序的潜能，包括让设计者说服人们参与其中，并保障他们隐私的一些重要技术。在一些案例中，鼓励人们参与的动机是明显提高的工作效率以及团队合作效率，并且提供给第三方的私人数据可以使用。在另一些案例中，设计者将博弈论的元

素融合到他们的应用中，让人们在参与中获得成就感。当参与其中的好处显而易见，而隐私又能很好地被保护时，人们就变得更加愿意分享个人数据，从而使他们的工作空间和社区对自己、对同事、对附近居民来说都变得更好。

社交网络的介入

Tacit是一家将自己的软件标榜为知识中介系统的公司，该公司的软件从员工发送出的电子邮件中搜集私人细节。这些细节，包括收件人信息、附件的内容、电子邮件本身的用词，被用来推断员工的社交网络、知识以及兴趣。因为Tacit使用的这种数据采集方法会让员工反感，所以该软件是基于严格并且透明的隐私控制规则来设计的。所有信息都被加密，这意味着公司内的任何人都不可能看到关于员工个人的兴趣以及社会联系的分析，即使是技术总监和CEO也不例外。同样，Tacit公司的其他任何人也都无法获取这些信息。仅在有法院指令的情况下，Tacit公司才会公布这些敏感数据，并且想要获取这些数据需要Tacit公司以及使用该系统的公司同时提供密钥。

那么，这具体是如何操作的呢？该软件只有在所有参与方都授权的情况下，才能显示关系信息。比如，假设罗杰认为他正在做的项目和公司的其他项目相关，但他不知道是哪几个项目，也不知道谁参与了那些项目。罗杰可以给不同的人发送电子邮件，描述他正在进行的项目并询问他们是否知道一些类似的项目信息；如果知道的话，他们可不可以向那些了解类似项目的人引荐罗杰。或者他也可以直接通过知识中介系统查询。该系统会持续地搜索与各种主题和项目分别相关的人，并私下询问符合条件的人是否愿意与罗杰结识，并和他讨论项目。尽管罗杰在启动搜索的时候表明了自己的身份，但只有接受了该

询问的人才会被真正介绍给他，其他没有接受询问的人仍然维持匿名的状态。

这一系统的理念是让分享知识、兴趣和建立社交关系变得更加便利，以提高工作场所的效率。帮助建立新社交关系的系统是传统知识管理系统的改进版，传统的知识管理系统一般主要用于文档的分析、编目以及分类，它通常距离实现真正的知识和关系创建仅一步之遥。像Tacit这样的自动系统，可以在人们需要使用专门知识以及人际关系时自动在系统中被发掘出来，同时，该系统还为那些不愿公开数据的人留有选择保密的权利。

麻省理工学院在2004年开展的Serendipity项目和Tacit系统有点相似，它们都是通过一个外部系统为人们提供中介服务。和麻省理工学院早先的“现实挖据”项目一样，Serendipity项目使用从参与者手机上采集到的数据，试图让拥有相同兴趣的陌生人在空间距离上相对接近的时候有相互认识的可能性。尽管，他们最开始的想法是在大公司或大型会议上使用Serendipity，将其作为交流的媒介，这个理念之后还催生了一家叫作MetroSpark的手机交友公司的产生。

MetroSpark公司要求用户在注册的时候提供关于自己的信息，并且按照重要性对不同兴趣进行权衡。根据这些信息以及兴趣衡量结果，系统会计算“相似分”。如果用户同意，MetroSpark公司会通过用户手机上的软件，对其行为进行推断来采集更多的间接信息，比如用户的睡眠时间、经常去的地方、社交关系，甚至手机上的游戏和其他应用程序的使用情况。另外，用户也可以选择他们想要结识的人的类型。所有这些信息组成用户的个人资料，并存储在中央服务器中。

用户可以自行设置相似分的临界值，当两个用户之间的距离在蓝牙传输的范围之内，并且如果他们的相似分都高于彼此自行设定的临界值，那么服务器会分别提醒这两个用户，在他们附近有个与自己兴趣相投的人。考虑到隐私和社交礼仪的问题，系统会通过发送短信提

醒用户，并询问他们是否有兴趣与这个人碰面。如果两个用户都选择了“是”，系统会把对方的照片以及一系列聊天话题发送到双方的手机上。

还有许多与MetroSpark相似的地理社交应用程序，比如，Grindr是一款以男同性恋为目标人群的移动约会应用，还有Blendr——一个目标是介绍具有不同人口统计特征（如职业、人种、文化程度等）的陌生人相互认识的手机应用。使用此类程序的用户可以上传并更新自己的照片和个人信息。以Blendr为例，用户可以扫描附近的用户信息（准确的位置信息是被隐藏的）并且标记他们感兴趣的用户。当一个用户标记了另一个用户，而对方同时也标记了他，那么这两个人就可以通过该程序直接交谈。

不幸的是，不是所有的软件开发人员在编写软件时，都考虑了用户的双向授权。我身边的女孩（Girls Around Me）就是一个不需要用户授权的应用程序，它很容易被匿名且有时是“掠夺性”的行为钻了空子。这个为iPhone设计的应用程序直接从脸谱网上链接个人公开信息，又从Foursquare上获取实时位置信息，并把这些信息都放在一个可以被轻易搜索到的地图上。在不需要任何授权的情况下，人们就可以通过这个应用程序搜索一定的区域，并看到那些将相关信息在线显示的用户，他们的照片、关系状态、兴趣以及精确的位置信息，但这些用户本身可能并不知道这些内容是对所有用户公开的。本质上，这个应用程序就是在用户不知情，且没有明确授权的情况下，组合用户的个人信息细节以及他们的具体地理位置。因此2012年苹果在线商店将之下线。Foursquare公司也引用“不允许整合不同站点的信息，以防对一系列地点信息进行不恰当的浏览”的政策条款，拒绝“我身边的女孩”访问其应用程序接口（aPi）。

脸谱网的隐私政策经常更新，而且使用起来也有很多不明确的地方，程序开发者们常常钻这个空子，从而导致很多用户的个人信息意

外泄露。相反，Metrospark、Tacit、Grindr和Blendr这些软件则需要用户的授权才可以分享个人信息，或提供中介介绍。然而，这些服务都存在于隐私的灰色地带，致使很多理智的用户觉得不舒服，因而不会选择使用它们。在这个领域，软件设计必须仔细考虑，而且在推广中保持透明也是至关重要的。

社区笔记的贡献

对于社区层面上的数据采集，地理位置信息附带上其他传感器所采集的信息，比如相机或加速计，即使只有一种，都有可能带来巨大的利益。当社区成员采集这些“协变量”并在地图上将其标记在特定的位置时，他们很可能发掘出自己社区中一些隐藏的特性，这些特性则蕴含着帮助改善所有人生活质量的潜在可能。这种类型的社区数据采集相比公司中的数据采集有一个明显的优势：你不需要说服所有人参与就能获得有用的信息，不像组织机构往往需要完整的社交图谱才是最理想的。

Noah项目让人们拍摄带有地理标签和时间戳记的植物、动物和菌类的照片，并将它们上传到一个在线生物体数据库。该数据库所在的网站提供了一个可供用户搜索的地图，而在该项目未来的更新中，将会允许人们创建他们自己的社区生态系统，以便学生和当地的野鸟观察家查阅当地的生态信息。Noah项目设置特定的任务吸引用户参与，这些任务一般是采集特定物种的照片和信息，在完成任务时，用户可以根据自己的贡献程度赚取类似功绩勋章的虚拟土地。

Noah项目将现实社区和在线社区连接在一起，让人们了解一定区域内一段时间中的生态系统的健康状况。这些数据现在被业余爱好者和伊利诺伊大学芝加哥分校以及康奈尔大学的科学家使用，分别作为

关于松鼠和瓢虫项目的补充数据。这些数据还可以用来标记濒危物种，追踪入侵物种，可以作为生态系统整体健康的指标。

Noah项目关注的是整个社区的生态健康，然而还有其他很多移动手机应用则是让用户报告社区中人工环境的生态健康状况。CitySourced,SeeClickFix,FixMyStreet以及其他很多移动手机应用都允许用户发送含有地理标签、时间戳记并且被预分类过的照片给市政府。在一些案例中，这些应用程序直接和城市后勤部门的工作流系统相连接，并且会自动生成工作指令或路线投诉发送给警察局，这意味着可以更快更有效地实现修补路面的坑洼或清除涂鸦。

让城市管理者、首席信息官以及市长支持这类计划并不困难。对市民有好处的对于政治来说也是好的。而这些应用程序中最成功的那几个，都是由那些致力于将他们的软件紧密整合到城市现存的基础设施中的公司所开发的，所以不需要在升级现有的信息技术上支出太多。

实际上，一些市政府是这方面的先驱，它们开始自行开发应用程序。在加利福尼亚州圣何塞市的"圣何塞311"系统是由CitySourced公司开发的，该系统让这个小公司迅速地发展起来。同样的，旧金山市将城市数据适当地对公众开放，包括街道清洁计划、停车信息还有再生设施信息等，用来鼓励所有可以提供有效分享城市信息方式的移动手机应用和网站的发展。

对于城市管理者和应用开发人员来说，他们现在之所以能够方便地合作，都要归功于一个名为Open311的初创企业，它是由纽约市的非营利组织OpenPlans开发的。其目标是建立一个开放平台，让城市可以利用现有的基础设施来发布城市信息，以便开发网站和移动应用。

在这一点上，大多数这类服务都允许用户匿名报告城市问题，服务器只能获取他们的位置。因为这项服务的收益——一个更清洁的社

区，只需要来自分散位置的个人用户提供的信息就可以实现，因此有很多的人愿意使用这些服务。然而这些应用程序的后续版本可能向用户注册并录入更多个人信息的选择。这样可以让用户更方便地获得针对他们所投诉问题的反馈，也可以在社区有活动时获得通知，还可以参与一些游戏和竞赛。

智能手机还可能通过另外一个途径来改善社区健康状况，即让人们与同一区域内的其他人共享自己跑步、骑行以及通勤的路径。从21世纪中期开始，有许多网站已经提供了可以根据位置和距离搜索的由用户生成的跑步和骑行路径。Mapmyrun.com, Mapmyride.com, Walkjogrun.net, bikely, Runkeeper, Runtastic和DailyMile.com都被广泛地用来分享和寻找路径。这些路径的数据既可以通过手动添加，也可以通过带有GPS芯片的设备上传。很多这类应用都可以让用户很容易地分享由iPhone和安卓设备实时跟踪记录下来的路线。

然而，可供搜索的自行车通勤路径还有一些未被开发的潜能，可以通过社区的众包来解决。众包就是一群人通过贡献自己的数据和资源来解决问题。自行车通勤者更愿意与其他考虑用自行车替代汽车通勤的人分享自己的窍门和策略，但现在还没有合适的方法来寻找从不同社区到市中心的可行自行车通勤路径，以及高峰时段这些路径的交通和道路情况（第三章中提到过UCLA的Biketastic项目，采集的就是这类数据）。

另外，对于新的自行车通勤者来说，预估在悠闲速度下的骑行时间（其中包括等信号灯的时间），是很有帮助的。像Runkeeper和DailyMile这样比较受欢迎的服务，一般都带有专门针对通勤者的搜索功能。应用开发者也可以开发能够自动采集自行车通勤者通勤数据的应用程序，并将数据上传到这些现有的网站上。考虑到隐私，Runkeeper和DailyMile都把个人活动路线的精确起点和终点隐去，让它们的用户在发布通勤路径时感到放心。尽管没有办法匿名发布路线，

通勤者并不需要经常与公众分享他们的通勤数据；一条通勤路径分享一次就可以给其他人提供有效的帮助。

挖掘社区中人们的行为，并把它们和其他诸如交通、空气污染或其他环境指标数据关联起来的工具，对于关心社区的积极分子来说会很有用。加利福尼亚大学和英特尔合作的“常识”项目（第三章中介绍过）尝试给人们配备带有蓝牙装置的空气质量感应器，感应器上的数据可以通过蓝牙传输到手机上。为了改善社区，积极分子可以向在公共空间活动的人们提供空气质量感应器，比如那些在公园游玩或在等公交车的人。如果这些感应器检测到空气污染物水平较高，那么就可以发起针对当地工厂的请愿运动，要求他们减少污染物的排放。

最近，旧金山一个著名的低收入社区Bayview开展了一个创新卫生保健项目，Bayview社区由于临近城市工业区，长期受到污染的困扰，这一项目希望改善该地区的压抑环境，减少可能导致哮喘、癌症、糖尿病和气胸等儿童及成人疾病的因素。这些应激原（即引起疾病的因素）包括较差的空气质量和噪声污染、虐待、健康食物的匮乏、锻炼不足，以及来自家庭和其他渠道社会支持的缺失等。

通过当地儿童的手机所记录的数据，包括他们上学的路径，等待公共交通工具的时长，距离贩卖新鲜食品的杂货店的距离等，可以有效识别贫穷社区的一些系统性的应激原。尽管儿科医生不能通过来自手机的数据解决所有问题，但他们可以给病童的父母提一些改变状况的建议，比如换条路去上学以减少遇到的汽车，或者尽可能地选择一条路过新鲜食物市场的路径。另外，移动数据还可以帮助卫生保健提供者们决定要为哪些社区提供新的娱乐设施和课后资源。这类干预措施即使没有来自手机的信息也显然是需要的，但具体的数据更有说服力，尤其是当卫生保健介入和卫生保健费用相关联的时候，这类方法会特别有优势。

对于这些项目来说，长期的激励机制可能包括兴建新的公园、娱乐中心，以及道路“减肥”，如减缓车速、拓宽人行道、种植绿色植物等。更直接和迅速的奖励机制可以是针对全部或部分参与者进行补贴，比如减免他们购买手机的费用或对话费进行补贴。因为家长完全有理由不信任那些监控自己孩子所处位置的技术，这个项目在整个社区中的接受度，有赖于那些有影响力的社区成员对这一技术及其目标进行宣传。

得益于合适的奖励机制以及设计精良的程序，越来越多的数据在小规模的群体中产生并得以采集，工作场所和社区中的政策设计也很可能从学术理论方法转换到实证方法。不管是为了促进小规模群体数据的合理使用，还是为了保护个人的隐私，都需要协议和规则；尽管如此，每个机构或社区的成员都有权从它们产生的数据中获得利益。如果应用开发者认真考虑如何鼓励用户参与和保护用户隐私，在小规模群体的数据采集领域将有很多不同的机会。

当考虑到工作场所的工作成果和效率，Tacit公司找到了一个最有效的方法来保证所有用户的隐私，即在用户和信息之间仅扮演中介的角色，只有在用户授权的情况下个人信息才会被分享。在通过手机报告社区衰败情况的案例中，人们似乎愿意向城市管理部门分享自己的匿名位置信息。游戏或其他的激励手段是否能让用户愿意亮出自己的真实身份，还有待观察；然而，越来越多的人开始使用位置应用Foursquare，而推特和脸谱网提供的数据也表明公众对于分享自己所在位置信息的抵触感也在下降。

当分享通勤、徒步或骑行路径的应用程序变得越来越受欢迎，类似Runkeeper和Dailymile这样的公司，需要仔细考虑它们的保障措施，以保证更多的用户能够安全分享数据。因为采集的数据并不多，而参与其中的用户是自愿将自己的数据分享到公共数据库中的，所以目前的隐私和安全措施在这个领域中比较适用。

然而，在污染追踪器以及医生向病人提供手机，以期找出社区中的应激原这两个案例中，隐私保护的局面发生了转变，道德问题随之产生。在这个案例中，可能需要依靠法律机制来说服参与者，他们分享的数据不会被用来针对他们，也不会提供给其他人。UCLA的卡蒂·希尔顿和德博拉·艾斯汀将这个问题明确地提了出来："如果原始的位置数据和经验取样数据在民事诉讼中可以很容易被获取，那么个人用户乃至所有民众可能都不会接受这种新的调研形式。"她们建议通过所谓的证据特权，来将个人数据等同于非商业交易机密来对待。通过这种特权，数据不能向特定的群体或机构公开，也不能被传唤或用于法律程序中。

直到现在，人们才开始认真地讨论关于个人数据的所有权性质及其保护。然而，即使没有这些保护措施，很多人仍然愿意与设计应用程序的公司分享大量的个人数据，比如位置追踪信息，甚至并不介意将这些数据公开发布。他们这么做是因为，他们觉得不可能有法律纠纷；并且对于他们来说，更好的信息能够获得更大的回报。因此，关于这类数据采集和分享的细微差别的标准和法律，可能需要很多年才能被大家接受。同时，最佳的实践方法是在所有应用中，简单并明确地告知人们，他们分享数据的风险以及好处。

---

1. "城市311"是纽约政府提供的一项便民计划，通过相关的电话、短信和网络渠道，人们可以获得政府信息，以及一系列非紧急性服务。——编者注

REALITY
MINING

Using Big Data
to Engineer a Better World

第三部分
大数据，让城市更美好

过去的10年里，很多犯罪数据都被电子化了，这就为数据挖掘提供了可能性，从中得到的模式可以帮助警察部门做出更好的决策：比如周五的晚上应该将警车派到哪些地段，或者像一场周日足球赛这种重大活动应该配置多少警力。

REALITY

MINING

Using Big Data

to engineer a Better World

第五章
城市数据的大用途

到2009年，全球70亿总人口中已有一半以上居住在城市。在美国，总人口的82%是城市居民，而印度则仅有30%。不管在哪个国家，城市都具有多样化的居民和行为特征，这些也都体现在城市数据中。本章讨论的对象是人口在1 000人到100万人之间的城市，并主要关注两种体现城市特征的数据：交通流量和犯罪率统计。

具体而言，本章探讨的是采集交通数据的多种途径，包括来自移动电话和车载GPS系统、道路探测设备的用户数据，以及来自传统的交警及其他交通服务机构的报告等，并对其中的隐私性和商业性的考虑进行了讨论。我们也注意到正在激增的公共犯罪数据库，警察、政府官员、居民或是房地产商都可以通过这些数据库查看城市中发生问题的时间和地方。本章还特别讨论了通过摄像头采集数据在法律和技

术上可能面临的挑战，这些问题在交通和犯罪监控领域都变得日益普遍。

尽管本章涉及的数据采集技术大多是简单明确的，但数据的内涵广泛。第六章考察了如何利用城市交通流量的数据来分配道路资源，并为驾驶者提供实时交通状况信息和路线选择建议。当犯罪数据与其他一些信息如天气、地理地形、重要事件、一年中的某些时间相匹配时警方可以利用这些信息，更有预见地进行犯罪监控。

城市空间层面的隐私问题与前两个空间层面有所不同。在城市中，人们有足够的理由希望保护自己的隐私信息。然而，如果发生了犯罪行为且嫌疑犯已经确定，那么这个人的匿名权就失效了。他的姓名、年龄、保释金等信息都会进入公共档案，并出现在某份当地报纸上。而且，警方还可能会向一些研究人员提供这些信息，以进行减少本地区犯罪的研究。本章也将涉及一些这样的案例。

交通数据

曾几何时，大多数的交通信息都是来自道路上方的直升机观测报告。如今，归功于GPS导航设备、交通摄像头，以及从驾驶员的手机和专为采集交通数据而设计的车载传感器中获得的位置信息，普通的驾驶者（以及交通报告员）可以获得比以往要详细得多的交通信息。而根据美国2011年的《城市交通状况报告》（Urban Mobility Report），交通拥堵每年给美国造成的经济损失已经超过了1 000亿美元，并且每年浪费每个通勤者长达34小时，因此有必要寻找一条更好的路线。

位于华盛顿州柯克兰的Inrix公司致力于通过更多的交通数据，为驾驶者的导航设备提供每分钟实时更新的路况。该公司是世界上最大

的交通数据供应商，这一市场的竞争者还有NACTEQ、TomTom International BV和Media Mobile几家公司。Intrix的交通数据服务的客户包括AT&T、宝马、福特、佳明、谷歌、MapQuest和Sprint等。

2004年，Intrix从微软公司独立出来，到2010年已经在美国实现了对260 000英里（约418 429千米）长道路的实时监测覆盖。Intrix用多样化的数据进行交通时间预测和路线建议，包括警方和紧急情况扫描摄像头提供的交通警示信息、各州交通部门（DOts）汇集的历史数据、道路传感器和摄像头的实时数据、越来越多驾驶者的手机和GPS设备的汇聚信息、重要活动和事件的信息，如演唱会或体育比赛，还有从车载设备采集数据中推断出的道路状况和天气。

传统的交通数据采集方法，基本上就是从地面或空中进行人工观测，且仍在使用，如无线电交通报告等方式。来自交警报告、紧急情况扫描摄像头，以及桥梁、隧道与高速公路管理组织和道路摄像头的数据被类似Clear Channel这样的公司汇总，并出售给广播、电视台以及Inrix公司等。研究人员也可以通过购买或授权获得数据，但这些数据一般需要用户特别通过设备制造商（如佳明和TomTom等）进行订阅而获得授权。

通过公共交通摄像头可以迅速且简单地判断一条道路是否拥堵，现在已有很多摄像头的实时画面可以通过互联网直接获取。这些摄像头通常被用于对交通状况进行定性评估，并快速发现交通事故发生地点。然而，也有些公司，如比利时的Traficon和以色列的Agent Vi等，正在生产可以自动分析视频数据的监测系统。

另一种监测交通状况的方式是通过道路植入式感应设备。这些被称为植入环路感应器的设备已经使用了数十年，并且还在提供有价值的数据。这种环路感应器可以产生一个电磁场，当汽车的金属底盘经过这个场时会对其产生干扰。到2009年，加利福尼亚州的高速公路系统下已经安装了25 000个这样的感应器，用于监测道路流量、拥堵情

况和速度。州和联邦交通部门保存了环路感应器采集到的历史数据，并向公众开放，有时也可以从交通部门的官方网站上获得。Inrix公司也将这些历史数据纳入了他们的交通预测算法中，基于这些历史交通数据，可以帮助人们设计交通线路，比如，提供一条在周四下午4点半穿越圣路易斯城区的路线。

还有一类感知技术是Inrix公司没有采用的，但可以用于进行与Inrix相似的交通流量估判，那就是固定式蓝牙探测器。马里兰大学2012年的一个项目表明，两个相距2英里（约合3.22千米）固定放置的蓝牙探测器可以精确获得交通流速度数据。目前，大约每20辆汽车中就有一辆安装了带有一个唯一识别码的车载蓝牙感应器，一般是用于连接手机以便在开车时不用手持手机便可通话。两个固定式蓝牙探测器中的一个捕捉到车载蓝牙设备的识别码，当车辆行驶到另一个探测器时，如果第二个探测器也捕捉到同一个识别码时，车辆的运行速度就可以被计算出来了。

尽管环路感应、蓝牙探测和道路摄像头是Inrix公司所需实时数据的来源，但这些技术并不能覆盖所有道路。这些基础设施感应器集中安装于主要的高速公路和域区。所以，Inrix公司开始寻求更加分散的交通数据来源点：车辆上的GPS导航设备和智能手机。虽然GPS导航设备相对较为精确，而手机的定位数据依赖于Wi–Fi和手机通信基站发射塔的三角测量，有时会有超过20英尺（约6米）的误差。不过，通过一定数量的手机数据，还是可以从误差中提取出有用信息。到2012年，Inrix公司已经获得了全球大约5 000万部手机和GPS设备的数据，大部分来自车队和商务车辆。该公司称，美国的众包数据主要来自车载GPS设备。2011年夏天，Inrix公司收购了一家利用手机通信基站三角测量技术对车辆进行定位的英国公司IT IS。尽管这项技术的精确度不如GPS，且在美国以外的地区只能从车流中获取大约50%的定位数据，但对于GPS设备不普遍的地区来说是一个很有用的补充。

Inrix公司也考虑到了城市及周边地区的重要活动，以及历史上类似活动的举行对交通状况的影响。举办活动的公司员工手动汇集活动的相关数据，之后这些信息被纳入交通预测和路线建议的算法中。

此外，Inrix公司还与一系列的汽车制造商合作，包括奥迪、尼桑和福特等，为它们的车载导航系统提供交通数据信息。2012年，Inrix公司开始从这些合作伙伴生产的车内传感器中挖掘信息。在众多传感器中，防抱死刹车系统和挡风玻璃雨刮器的使用情况也成为判断道路实时状况的重要信息。

与Inrix公司一样，谷歌也使用手机帮助判断道路交通状况。除了使用车队数据外，谷歌还可以从那些允许谷歌地图读取GPS信息的用户手机上，以匿名方式获取定位和速度数据。据谷歌公司的说法，用户的手机会匿名向谷歌发送手机移动速度的数据。结合同一时间道路上其他用户的手机发送来的速度数据（如今超过2亿部移动设备安装了谷歌地图），这种方法提供了非常不错的实时信息。但是谷歌缺乏有些国家的数据，所以从2011年秋开始，它们与Inrix公司合作来填补这方面的空缺。

谷歌考虑到了用户的隐私问题，它们采集的速度和位置信息是匿名的，并且需要用户同意才能上传数据至谷歌的服务器。此外，当多人从同一区域发送数据报告时，谷歌会将这些数据混合起来，从而很难去区分不同手机的上传内容。最重要的是，每一段行程的起点和终点数据会被永久性抹掉，即使是谷歌员工也无法获知相关信息。而在手机上允许谷歌地图自动上传数据至服务器的用户，也可以选择通过禁止定位服务来停止上传数据。尽管谷歌为其一些地图数据提供了应用程序编程接口，包括方向、距离和海拔，但是目前交通数据对于那些想要开发它们自己应用的程序员来说依然是不可得的。

2013年中期，谷歌收购了硅谷的创业公司Waze，该公司的业务主要是在手机用户允许的前提下采集交通数据。实际上，2009年成立的

Waze公司主要依靠用户自行充实地图内容（比如，补充空白小路的路名），还通过用户上报车辆测速区、交通拥堵和事故。2012年，Waze公司估计其用户数量比先前翻了一番，从1 000万涨到了2 000万。随着更多城市的更多用户参与进来，Waze公司可以提供更准确的行程时间和更好的路线选择，它们的服务因此也变得更加有价值。为了吸引更多用户的参与，这家公司还引入了“社交游戏”元素，向使用软件的用户发放虚拟奖励、积点和徽章。

谷歌、Waze和其他一些公司开发的这类手机交通应用，能够吸引用户的原因主要是便宜且提供比较可靠的交通预测数据，并且还有一定的使用乐趣。虽然Inrix和TomTom公司也提供手机应用，但它们的核心业务是为车载导航系统提供精确的交通分析数据。目前，Inrix和大众来源数据软件的交通数据还无法免费下载，尽管有些可以通过购买使用许可的方式获得。2008年以来，马里兰大学与Inrix公司和I–95廊道联合组织[注]建立了合作关系，研究人员通过利用Inrix公司在I–95交通廊道（贯穿美国整个东海岸线，北起缅因州南至佛罗里达州）沿线10个州，超过20 000英里（约32 000千米）道路上，采集汇总的实时和历史交通数据，来帮助美国交通部研究更好地分配交通资源及分配地点的方法。

用数据预测犯罪

预测犯罪行为的关键是犯罪历史记录的基础数据库。如今最可靠的数据采集手段与过去相比并没有什么不同：警察仍然需要写事故和逮捕报告，其中包括了很多有用的信息。这些案件报告主要有两种来源：一是警察在巡逻时发现犯罪行为，对嫌疑人进行盘问并决定是否逮捕；二是有人报警提供消息。警察会在报告中记录时间、日期、犯罪行为发生地点和类型（有数百种类型），还有更具体的信息，如周

边环境（比如是在便利店里、小路上或是街道上），如果是有人报警，还会记录报警人的姓名和联系信息。

数十年来，事件和逮捕报告数据都被输入电脑进行处理并标注于地图上，形成更有用的视觉空间格式信息。这些地图一般被用于一些特殊案件，例如用于界定犯罪嫌疑人活动的地理范围，还可以从中分析特定社区可能的多发犯罪类型，以及随着时间变化，犯罪行为向新社区的空间迁移和扩散情况。

近年来，专业算法和实时犯罪数据的使用可以对这些犯罪地图进行持续升级，从而可以在犯罪发生之前先将警力部署至事发地（详见第六章）。在田纳西州的孟菲斯市，警方从2005年就开始使用一款叫作Blue CRUSH的软件。通过历史犯罪数据的热点地图，警官们可以检查当前的犯罪活跃度，以及因为之前的警力部署而导致的相应区域的犯罪水平变化情况。每周更新的热点地图会被用于制订下一周的人员部署策略。

除了能生成显示犯罪行为的空间关系图，犯罪数据还可以与多层地理信息图相叠加，如美国地质勘探局提供的地形数据。另外，也可以将罪犯和受害人的住址标注在地图上并叠加地理信息，如道路、学校、选区、铁路、工厂、就业岗位、平均收入等可以从美国统计局免费获取的各种数据。同样，不同犯罪数据之间还可以进行比较，以获得更深入的了解——比如查获毒品和汽车盗窃发生时间的比较。

除了绘制犯罪信息地图，一些研究人员和警力部门也在编辑可疑帮派活动网络图。为了弄清楚帮派之间的活动关系，洛杉矶警察署和加州大学洛杉矶分校分析了超过1 000起帮派犯罪的历史数据，以及10年间对于一个拥有大约30个帮派地区的犯罪嫌疑活动数据的记录。研究人员可以通过模型来判断最有可能进行新犯罪活动的前三个帮派，准确度达80%。

圣克鲁兹大学的研究人员采用了加州大学洛杉矶分校的另一个模型，通过圣克鲁兹警察部门的数据来监控犯罪行为的爆发。他们的研究发现，犯罪行为的发生与地震后余震的发生具有类似的数据模式，该发现指导警方在最有可能发生犯罪的区域部署人员。

用视频监控犯罪

在街道上安装视频监控摄像头是一种数据密集型的犯罪信息采集手段。美国国土安全部为了对抗恐怖主义的威胁，为警察部门安装摄像头提供了大量资助，这使得美国运用于公共安全方面的摄像头数量出现了井喷。2009年，国土安全部为其城市安全项目（Urban Area Security Initiative）在7个城市花费了1 500万美元，该项目的目的是为了在城市地区，针对恐怖主义行为的防范、应对和重建问题，进行规划部署和培训。2010年，该项目在64个大都市区花费了8.3亿美元。2011年，31个城市投入共使用了6.62亿美元的项目资金。

关于这些摄像头是否能够对犯罪行为起到震慑作用，则有许多不同的观点。南加州大学2008年进行的一项研究，通过对洛杉矶犯罪视频的分析发现，从统计上来看，摄像头并不能显著地降低暴力和财产犯罪的发生。同样的，加州大学伯克利分校在2008年的一项研究中也认为，旧金山在安装摄像头之后，暴力犯罪的发生并未减少。伦敦以其众多公共摄像头而闻名，2009年来自伦敦警方的一份内部报告则称，摄像头对于帮助抓捕罪犯并没有什么帮助。

但城市研究所2011年发布的一份研究报告，通过分析芝加哥、巴尔的摩和华盛顿特区2001年以来的历史数据，揭示了更加复杂的结论。芝加哥的监控摄像头网络在这三个城市中覆盖最广。洪堡公园地区安装了摄像头后，总体犯罪率下降了12%；然而芝加哥的另一个地区，西加菲尔德公园地区的犯罪率没有变化。巴尔的摩市的三个被研

究地区中，一个地区的犯罪率下降了25%，另一个下降了10%，剩下一个地区没有下降。在华盛顿特区则没有出现犯罪率下降。该研究报告的作者认为，对摄像头技术的最佳使用方式是，让训练有素的人员监控实时屏幕，能够移动摄像头以获得最大视野。这一建议点明了摄像机装置以及维护此类数据采集的重要性。

如何获得公众数据

尽管对普通市民而言，他们并没有获取公共摄像头大部分内容的权限，但市民可以通过诸如crimereports.com或是crimemapping.com这样的网站连接到交通摄像头、网络摄像头以及匿名的犯罪行为数据库，这些网站依靠执法部门提供犯罪数据。用户都可以通过这两个网站看到不同犯罪行为在地图上的可视化分布图。为了保护犯罪受害人的隐私，具体地址被隐藏，所以案件发生地的数据是街区层面的。数据不能被下载，且通过技术手段抓取数据也是违犯用户使用条款的。然而，在crimereports.com的网站上，性侵犯案件的详细信息如姓名、年龄和地址都是可以查看的。

监控的合法性讨论

当交通摄像头被安装在距司机有一定距离的位置时，看起来似乎是无害的，但2004年美国民权同盟就此抗议，称警方可以使用这些实施监控的交通视频数据非法拦截车辆。该组织认为这违犯了禁止非法追踪和抓捕的禁令。

近年来，开始出现一种新类型的道路摄像头。除了监控交通流量外，它们还能抓拍驾驶员面部和汽车牌照，从而据此对闯红灯和超速

行为开出罚单。但2007年明尼苏达最高法院认为在红灯时对驾车者进行拍照，侵犯了司机的无罪推定权 (注) 。其他一些州，包括威斯康星、西弗吉尼亚、南加利佛尼亚、新罕布什尔、蒙大拿和密西西比等州都禁止使用这种摄像头。而部分大一些的城市如芝加哥、巴尔的摩、圣地亚哥、波特兰和华盛顿特区等，还在继续使用红灯摄像头，这些设备通常都是由私营企业安装和维护。民意调查显示，这些城市有超过半数的民众赞同使用这类摄像头。

用于交通和犯罪监控的公共摄像，或统称为静默视频监控，在法庭总体上还是得到支持的。1967年美国最高法院在联邦宪法第四修正案的基础上，通过“卡茨诉美国政府”案（美国最高法院判例汇编索引号：389 U.S. 347）对现代意义的搜查与扣押进行了定义。基本上，在公共道路上行走或站在公园中的人，其活动是不被认为具有隐私权的。同样，在公共道路上驾车行驶的人也不能要求隐私，他是可以被监视的。因而，在道路上对个人进行公共监控就是可行的。

而在1993年的“美国政府诉谢尔曼”案（索引号：990 F.2d 1265）中，这类监控的合法性得到进一步强化。第九巡回上诉法庭认为，在公共空间被摄像的人“不应期待获得隐私权，也不能以违犯第四修正案为由起诉政府进行影像记录的行为”。

在公共场所使用摄像头的合法性表明，研究人员也可以在公共空间自行安装摄像头进行相关研究。实际上，之前已经有在大学校园安装摄像头进行人员流动性研究的先例。然而自行安装摄像头的花费颇大，很多在项目中使用摄像头的研究人员，都是使用提供在线影像的政府资助和运作的摄像头。目前，利用网络摄像头的研究，大多数是依靠远程设置的摄像头追踪野生动物或进行环境观测。

虽然对于个体研究人员或企业主来说，想要获取整个城市的交通和犯罪数据还是相对困难的，但是他们可以选择与那些采集和汇总数

据的大公司合作，或者与警方合作以获得进入历史数据库的权利，或利用可以公开下载的政府数据。当犯罪数据被小心谨慎地使用时，不仅可以避免侵犯个人隐私，还能被用来保护城市安全。

此外，研究人员还可以在特定地点安装（非音频的）摄像头，自行采集城市流动性数据，尽管数据样本规模较小。但即使是这些关于某一部分城市生活的小样本，经过时间的积累，仍然可以提供对城市的深入认识。警力和交通资源可以据此在城市中得到更好的配置，一些研究人员甚至还开始利用这类数据监控疾病的爆发。目前，对犯罪和交通数据的利用还很不足，它们其实可以让城市变得更加安全和宜居。在下一章中，我们将探索在这些数据中发现的一些可能性。

---

1. I–95是美国95号州际公路的缩写，全长约3 098千米。共计跨越15州，是美国东岸的交通动脉。——译者注

2. 无罪推定权，指任何人在未经依法判决有罪之前，应视其无罪，被告人不负有证明自己无罪的义务，被告人有提供证明和有利于自己的证据的权力。——编者注

第六章
将适合的资源放在适合的位置

随着技术的发展，人们对数据的分析能力也从个人层面扩展到整个城市范围。有些人对待数据集的态度，就如算命巫师面对水晶球一样：似乎提出准确的问题，未来就会展现在眼前。然而，即便拥有了大数据和有效的分析手段，未来的画面也不会那么清晰。不过，城市规模的分析给我们提供了一个令人激动的机遇：为犯罪和交通问题建立快速适应系统，这样你的预测就可以在这些领域发挥一定的作用。

通过分析犯罪数据中的趋势，费城、孟菲斯和洛杉矶的警察部门在此基础上建立了一个能够在案发前对警力资源进行最优分配的系统。这些系统分析的结果表明，在特定区域只要有警察出现就有可能阻止严重犯罪的发生。当在高犯罪地区以一种不威胁那些遵纪守法的市民们的、温和的方式布置一定警力，这些犯罪预测系统就可以让社区变得更加健康而且安全。

交通预测是另一个有发展前景的领域。像Inrix和谷歌这样的公司将众多的数据来源整合起来，以更好地预测出行时间，同时也不断更新实时数据，帮助人们避免拥堵或事故。微软公司的研究人员甚至可以通过交通数据，预测未来的意外交通事件。正是因为有了现实挖掘，我们才可以预测意想不到的事件。

但是交通数据并不仅仅只能用来预测交通状况。城市规划师们也可以用这些数据来分析哪些道路和交叉口存在危险需要升级，如何在紧急状况下最快疏散人群等问题。此外，交通数据也可以用来生成城市及周边整体机动性的鸟瞰视图，让我们看到城市居民的暂时性流动。

在理解城市机动性的基础上，研究者们可以针对疾病传播提出一些重要的问题：某个病毒是在何处起源的？哪些人需要被隔离？哪些人需要被注射疫苗？一些研究人员使用交通数据和其他一些机动性指标，来更好地理解传染病在城市间的传播（第十章中会更加详细地讨论在全球范围内对疾病发展的追踪）。

本章主要讨论在城市范围进行的几项更细致入微、简单预测性的现实挖掘应用。

交通预测和意外事件控制

Inrix公司编写了一套基于多种来源数据的算法——包括手机GPS和无线定位数据、车载GPS的位置信息、道路感应器和摄像头捕捉到的信息、历史交通数据、重要活动安排和天气预报信息，综合利用这些数据来判断每一段道路的交通流速度。然后，该公司依据这些信息来计算和估算整个行程的驾驶时间。此外，这个算法还通过持续监测交通流，不断更新估算预测结果。根据马里兰大学和I–95廊道联合组织共同领导的一项独立研究——他们在I–95交通廊道上的一段2英里的道路上设置了固定式蓝牙感应器，以进行交通流监测（详见第五章），监测发现Inrix公司的速度预测精确度可达每小时2英里且预测同期为一星期7天，每天24小时不中断。

2008年，谷歌公司为网页版的谷歌地图产品开发了一个新功能——预测用户未来一段行程所需要的驾驶时间。如果你想知道去见医生开车需要多久，只要输入约见医生的日期和时间，谷歌地图就可以为你计算和预测。但是这个功能很快被取消了，并且至今也没有重新恢复。需要指出的是，现在苹果和安卓手机上的谷歌地图应用都可以估算驾驶时间，但是并没有可以估算未来旅行的驾驶时间的功能。

尽管这些交通预测系统已经相当完善，但在意外情况发生时它们还是可能会犯错。比如一辆运输鸡蛋的货车侧翻，或是一次临时的道路管制等，这样的道路意外事故的发生几乎无法预测。但是，微软公司的研究人员探索出了一种方式来预测什么时候有可能发生这些意外，他们称之为“意外事件模型”。这项技术可以用于提升交通预测的精确度，也可以用于其他一些需要预测的领域，比如健康保健、军事战略以及金融市场等。

意外事件模型是微软一个名为SmartPhlow的软件开发项目的成果，该项目是由埃里克·霍维茨和他的同事们于2003年发起的。这个软件可以展示公路上的交通流状况，还可以在发生意外事件时提醒用户，比如当一条道路被临时管制，或是一条畅通道路突然变得拥堵。研究人员通过采集使用西雅图多年的交通数据，将事故、天气、节日和活动数据与之进行关联分析，并将每天的时间切分为多个时长为15分钟的片段，对每个片段的交通状况分别进行计算。他们特别关注数据与一般模式有显著差异的情况，即异常的交通状况的出现。在对这些意外事件进行标注之后，（研究人员发现）异常情况发生前30分钟内的事件都有可能引发交通异常数据的出现。然后，研究人员检查了这些意外事件的先决条件来研究其中是否存在一定的关系模式。该项研究成果是一个可以预测大约50%的道路意外事故，但会有5%的错误率的软件系统。

道路资源配置

I–95是一条连接马里兰和佛罗里达的南北向公路。I–95廊道联合组织作为一个合作机构，包括了交通部门，东海岸各州公共安全组织、港口和交通运输组织，马里兰大学以及Inrix公司。自2008年，I–95联合组织已经开始研究如何将海量的交通数据转变为可以改善交通

网状况的有用信息，并开展了一个基于车载GPS等公众数据的，名为VPP（车辆探索项目）的新项目，并计划将项目至少延长至2014年。

这个项目的目标是通过诸如I95travelinfo.net和511一类的网站和电话服务，为驾驶者提供更加准确的信息，并且帮助城市规划机构更好地进行决策。比如，华盛顿州、华盛顿特区以及费城的规划师就在使用VPP项目数据来评价目前公路网的效率。这是通过规划改善公路状况的第一步，从而最终减少道路拥堵并改变整体交通模式。

VPP项目的大量数据还可以帮助紧急响应团队更快到达事故现场。例如，2008年10月一次突发的暴风雪中，新泽西的交通管理员通过VPP数据，发现了80号州际公路的延伸段上的一系列事故，从而快速地派出了应急处理团队，这在原本基于摄像头的监控系统中是做不到的。

县和州政府也可以使用交通数据在灾难发生时更快疏散民众。佛罗里达州是个很好的例子，这里每年都会有数次飓风光顾，为保证灾害期间交通疏散得平稳流畅，该州方面主要是在县政府层面进行协调，另一方面与州政府的各部门在更高层面进行协调。

佛罗里达州的政府部门通过实时交通数据，可以获知灾害期间市民的驾车行动方向，并通知邻近的县和州，以保障足够的资源。此外，在特别拥堵的疏散路段可以实施“逆行”的方式，引导疏散人群向其他疏散路线或紧急避难所方向前进。

可追踪的病菌

大量交通数据带来的另一个好处是，我们可以以鸟瞰视角观察城市中人们的流动方式，这对于流行病学家来说极具价值。研究人员如

今通过交通数据，来追踪人们以及他们可能携带的病菌是如何在城市中移动的。

城市流动性数据可以用于疾病传播模型，以评估疾病是从何处发源的？将传播到哪些地方？以及将以什么样的速度传播？这些信息可以进一步被用于制订合理的隔离和接种疫苗方案，以及将医疗人员等资源在正确的时间配置到正确的地点。

包括亚历山德罗·韦斯皮尼亚尼和德克·布罗克曼在内的许多研究者，都利用交通数据对不同空间范围上的疾病传播网络中的流动性进行建模。韦斯皮尼亚尼和他的同事们发现，城市范围的地面通勤与航空运输的交通流有所不同。城市范围的通勤交通与空中交通相比，具有数量级上的强度更高，这表明城市范围内人们有越来越多样化的互动和联系。这一发现强调了理解不同范围流动性数据之间细微差别的重要性。韦斯皮尼亚尼和他的同事们提出，全球疾病网络模型中应该包括城市范围的模型，从而使总体模型更加精确并提供多层次的粒度。

预防犯罪

警察部门当然有采集犯罪数据的职责：案件报告中需要包括非法嫌疑活动的人物、内容、事件、地点以及发生原因等记录。过去的10年里，很多这类犯罪数据都被电子化了，这就为数据挖掘提供了可能性，从中得到的模式可以帮助警察部门做出更好的决策：比如周五的晚上应该将警车派到哪些地段，或者像一场周日足球赛这种重大活动应该配置多少警力。

纽约市的数字化犯罪统计系统Compstat是个很好的例子，它从1994年开始投入使用，如今已经在全美多个城市中得到应用。

Compstat主要追踪被记录的非法活动中的长期趋势。富兰克林·齐姆林在他2011年出版的著作《更加安全的城市》中提出，Compstat在过去几十年里对纽约犯罪率的下降具有积极的影响。通过Compstat系统，警察部门可以在预计可能发生犯罪的地方配置更多的警力。

在过去的10年中，很多警察部门都受到了Compstat系统的影响，但也推动了该系统的新进展。然而，Compstat数据一直只能用于每周的战略部署讨论，而且需要依靠人们来识别数据中的模式。现在有一些新的犯罪追踪系统可以每日更新，并依靠算法几乎实时地预测未来可能的犯罪。

2004年，来自卡内基–梅隆大学的研究人员与匹兹堡警察部门合作，开展了实时犯罪追踪的很多基础研究工作。杰奎琳·科恩、威鹏·戈尔以及安德烈亚斯·奥利希施勒格尔梳理了匹兹堡市从1991年1月到1998年12月间的130万份案件报告记录的犯罪数据，包括从盗窃到谋杀的16种犯罪类型。研究者希望能提前一个月预测城市中可能发生的犯罪活动，并在一个划分为104个地块单元的网络系统中预测其可能的位置，这些单元每个覆盖大约在100个街区的范围。

研究人员发现，他们可以找到与严重犯罪活动相关的其他犯罪指标，并确定这些指标的影响程度。利用这些信息，他们开发了一个预测模型，帮助警方决定应该向哪里分派警力，哪里的警力应该被撤回。因为人们普遍认为，警察的存在可以有效预防更多犯罪的发生。这个模型主要依靠犯罪活动之间的相关性进行预测。研究人员认为，传统方法仅仅根据单一犯罪类型进行趋势分析，而他们的相关犯罪模型与传统推断方法相比具有显著提升。

虽然匹兹堡和纽约的犯罪率是有所下降的，但在圣克鲁兹市却在上升。2000~2011年，该市的犯罪数量上升，而更不幸的是，由于州预算和城市预算的削减，警察数量反而在减少。2011年，圣克鲁兹市的警察部门开始与乔治·蒂塔、乔治·莫勒、马丁·肖特，以及杰夫·布兰廷

汉姆等研究人员合作，开发并测试了一个有效的犯罪预测系统，以协助警察安排更高效的日常巡逻。研究人员将城市划分为500英尺（约152米）见方的网格，并标注上过去8年圣克鲁兹的案件报告数据。与纽约的Compstat系统不同，这个系统每日更新数据。并且，它还采用了以前用于预测地震余震的计算机模型，并将其很好地应用于预测犯罪活动，警方可以通过这个系统得到每日更新的犯罪热点地区。据说，这个系统或多或少降低了犯罪发生的频率。然而，在本书写作时（2013年），还没有最终的分析结论。

与圣克鲁兹一样，孟菲斯市在21世纪也经历了犯罪率的上升。但由于政府工资冻结和预算紧缩，没法雇用更多的警察来覆盖更多的城市范围，所以，在2005年孟菲斯警察部门开始与IBM公司合作，开发了第五章中提到的“Blue CRUSH”系统。这个系统使用了过去10年中的案件报告数据，以及现在警察的手持数字设备中的实时更新数据。基于这些数据，系统建立了犯罪活动、位置和其他一些变量，如废弃房屋之间的相关性。除了协助警方安排巡逻地区外，这个系统还能帮助更加有效地安装安全监控摄像头，因为它可以分析在哪些地点需要安装摄像头，哪些时间需要特别监控。在使用Blue CRUSH系统后，孟菲斯市的严重犯罪活动数量总体下降了30%，暴力犯罪则减少了15%。

很明显，还有很多不同的方法和模型可以用于犯罪数据分析，它们的选择取决于警察部门需要解决什么类型问题，以及需要配置哪些资源。除了可以减少犯罪活动的发生，这些系统还可以协助城市政府部门更加合理有效地分配警力资源，特别是在圣克鲁兹和孟菲斯这种城市预算有限的情况下。

交通和犯罪数据都很适用于建模和预测。帮助建立更好的疏散人群、应对紧急情况以及追踪传染病的扩散机制，只是交通数据的少数几种用处。突发事件建模，最初被用于预测意外交通状况，之后也被应用于犯罪事件等其他领域。

以何种方式来应用犯罪预测系统，以及该系统将如何影响高犯罪率社区，是开发该系统时需要特别注意的两点。对于警方来说，根据直觉和工作经验安排驾车或步行的巡逻路线，并前往潜在犯罪居民区或大楼巡视，并不是一件新鲜事。但是，安排警力在特定地点守株待兔，等待事件发生就是新鲜事了。如果没有恰当的培训和监管，这些系统可能会被用来对市民进行定义和分类，从而可能打扰普通的市民，甚至导致不必要的抓捕。犯罪预测系统的作用和潜力令人印象深刻，这些系统必然获得更多关注。而当它们被广泛使用时，如何合理运用系统尤为重要。或许在这些新技术全面铺开之前，可以让社区负责人参与到这些技术使用的讨论中来。这样做不仅能听到普通民众的忧虑并建立彼此的信任，还能借助基层居民的深刻见解提高犯罪数据的真实度。通过检查非法活动数据在现实中的真实状况，建立一个相互尊重的沟通方式，将是保证犯罪预测系统项目整体成功运行的有效方式。

REALITY MINING

Using Big Data to Engineer a Better World

第四部分 大数据治国

当现实挖掘的规模继续扩大时，国家政府、大型企业和国际组织就开始在数据的采集、编辑和传播上扮演重要的角色。在国家层面上，研究人员和企业可以获得更大范围的数据来源，包括国家人口普查、通话记录、主要互联网公司如谷歌、脸谱网、推特等，以及有限的一些银行数据。

REALITY

MINING

Using Big Data

to engineer a Better World

第七章
当数据上升至国家层面

当现实挖掘的规模继续扩大时，国家政府、大型企业和国际组织就开始在数据的采集、编辑和传播上扮演重要的角色。在国家层面上，研究人员和企业可以获得更大范围的数据来源，包括国家人口普查、通话记录、主要互联网公司如谷歌、脸谱网、推特等，以及有限的一些银行数据。当然，这些数据来源中有些相对比较容易获得，有些则相对困难。

到目前为止，人口普查数据是最容易获取的。很多国家把他们的人口普查结果通过网络向公众公布，这些数据可以被下载、可视化以用于更深入的研究。另外，世界银行开展的国际研究将所有参与国的人口普查数据汇编在一起，对其所有成员国来说就像一个一站式的信

息商店。这些数据是完全公开的，人们可以下载并被独立地分类和分析。重要的是，世界银行提供了开放的应用程序编程接口，让软件开发者可以将各种数据整合到应用软件中。谷歌就将世界银行的数据整合进一个简单的可视化工具，显示在其搜索结果中，比如搜索博茨瓦纳的人口时，将会显示世界银行提供的历史数据，以及展示人口数十年来变化的图表。

另外一个新兴的数据来源是通话数据记录，或通话细节记录，它们统一缩写为CDRs。CDRs数据特别有助于了解人口在一个国家或一个区域内的流动性。然而，与人口普查数据和世界银行数据所不同的是，CDRs数据对于普通的企业和研究人员来说很难获取。

作为一个数据集，一条CDR中包含了通信记录（通话和短信）以及事务性事件，包括呼叫方或发送方、接收方以及时间、地点、通话时长等信息。以往，CDRs数据仅被用于账单事务，但2005年开始，网络服务提供商和大学的研究人员开始意识到这些数据的价值，特别是对于人口流动性的建模。一些研究人员、企业与移动运营商签订了协议，可以在一定限制条件下使用他们的数据。一些移动运营商愿意分享匿名的CDRs数据，只要法定协议中详细规定不可以公开所有权以及个人信息。这些协议的附加条款往往会规定研究人员需要向运营商展示他们打算做的研究的价值，比如，一个预测性“搅动”模型的开发，涉及的是终止订阅和产品接纳情况。

在国家层面上，我们也同样强调主要网络数据采集商——谷歌、脸谱网和推特的重要性。这些公司对使用它们的个人、社区以及政府都产生了深远的影响，但它们作为大量数据采集工具的功能，在国家层面上变得更为明显。互联网公司采集数据的一个显著应用就是定向广告。但这些数据还有很多未开发的可能性，包括定向的市场调研和疾病追踪（详见第十章）。推特的设计理念，是让大多数用户所输入

的内容能够实时公开显示，它可以被用来追踪民众的情绪，也可以被用来追踪国家灾难或者其他危机发生时的资源分配。

一些关于谷歌、脸谱网和推特的应用会在第八章中详细讨论，还有一些将在第十章中突出强调，因为这些国际公司在全球层面上也有很大的影响。实际上，本书的第四部分和第五部分提到的一些数据类型，比如CDRs数据、网络公司拥有的数据集、银行数据等，都可以在国家层面和国际层面采集，并在之后被应用于这两个规模层面。本章初步展示了可以在国家层面采集的一些数据，并讨论了如何处理这些数据。我们同样关注隐私问题，特别是在匿名CDRs数据的应用上。尽管匿名记录上的个人识别信息被剥离，但通过与一些诸如人口普查数据等容易公开获取的信息进行交叉分析，还是有可能还原个人识别信息的。我们介绍了目前正在开发的一种很有潜力的方法，可以让研究人员和企业在获取CDRs数据的同时保持它们的匿名性。

人口普查

当今世界的大多数国家都有人口普查，其中多数国家通过网络公开发布这些数据。例如，美国政府通过一个名为美国事实查找器（American Fact Finder）(注) 的网站公布人口普查数据。人们可以在该网站通过主题（如年龄和收入），地理（如州和郡），种族、民族或部落，产业、产品或商品名称进行搜索。在人口普查数据之外，人们还可以下载其他的研究数据，包括每5年开展一次的经济普查，还有人口预测项目，该项目公布最新的各乡镇、城市、州和国家的人口预测数据。

2009年，美国政府启动了一个名为data.gov的在线数据库，提供来自各个政府机构的超过445 000条原始数据和地理空间数据集，所含门

类包含商业、艺术、娱乐和出行、选举、对外援助、交通、福利和其他；这些数据集有“性别/种族发展趋势”“美国的活跃矿区和矿井”“野马和野驴数量统计”等。这些数据可以XMl、文本或CSV、KMl/KMZ、Feeds、XlS和ESRI的Shapefile格式下载。此外，还有一些常用的交互数据集拥有自己的应用程序编程接口，开发人员能够把数据生成的地图和图表整合到网页应用中。

世界银行是国家层面公共数据的一个巨大的来源。这个国际机构提供的数据目录包括数据库、预定义格式的表格、报告以及其他资源。总体上，该网站提供超过7 000个指标的下载，涵盖了从农业土地比例到受教育比例等各方面数据。另外，该网站提供超过200个国家和经济体的数据，从主题上分有健康、气候变化等，从指标上分则包括国内生产总值、汽油价格等。

世界银行发展数据中心[注]整合了数据采集和统计分析，并维护世界银行，用于制订援助计划、评估贫困程度以及开展研究的一系列数据库。大部分数据来自187个成员国，它们和世界银行及其合作伙伴共同致力于改善国家统计系统的能力、效率和效力。

世界银行宣称，它们采用的是目前可能的，最专业的数据标准来运营，使用两大主要框架来评估数据统计系统：一般数据发布系统和数据质量评估框架，两者都是和国际货币基金组织共同研发的。然而，世界银行也承认数据集也存在不一致现象，因为不同的参与国数据统计的时间和编制报告的方法不同：所有数据都是通过传统调研的方式获得，从而导致数据集的数量相对较少。因此，世界银行建议在组合数据集的时候要多加小心。就其本身而言，世界银行提供了一套整合经济数据、增长率计算以及受汇率影响的换算系数的方法，这使得跨数据集的比较尽可能的精确。

世界银行的网站除了搜索数据集和下载表格外，用户还可以通过三个不同的应用程序编程接口来接入数据：一个接入指标，一个接入项目（世界银行运营的数据），还有一个接入世界银行的金融数据。这些接口使应用程序可以在可调节参量范围内通过子查询引用超过3 000个指标。世界银行提供的应用程序编程接口可以获取近50年间的许多数据集，并包括了已经结束的项目、活跃的项目和正在进行的项目的数据。

谷歌在一个名为谷歌公共数据管理器的可视化和比较工具中整合了世界银行的数据，让用户可以对来自世界银行的数据和其他60个不同公共数据集的数据进行交叉分析。然而，要进行更复杂的分析的话，用户必须单独访问每个数据集各自的来源网站。

通话记录

为了给客户生成账单并了解其网络上的通话负载，移动手机网络服务商会记录它们用户的手机使用情况。CDRs数据中包含了一个通话或一段短消息发送的时间、大概位置（通过通信基站的编号）及其接收者等信息。因为当今世界手机高度普及，CDRs数据对于人口流动性研究来说是非常有吸引力的位置信息来源。另外，CDRs数据提供的位置信息价格便宜，尤其是与需要耗费大量人力的小规模调研相比。

近年来，网络服务供应商、大学以及初创公司的研究人员开始注意到这种类型的数据对人口流动性和社交网络研究模型的影响（第八章介绍了基于CDR模型的特定应用程序）。然而，对于大多数研究人员和企业来说，获取这类CDRs数据是一个极大的挑战。出于隐私考虑并依照所有权规定，移动运营商不能对公众公开它们的数据。因此，如果一个研究人员想要获取这些数据，他需要能够提供特定的研究技术来改善模型，或是提供对于运营商来说有价值的其他东西，比如评

价一个客户终止他的移动服务的可能性。如果研究人员或企业进行CDRs数据研究的目的不能直接满足移动运营商的需求，那么他们可能需要额外花点研究时间为运营商提供直接援助。

移动运营商在分享这些数据上犹豫原因，部分来自与公众相关的隐私考虑。而且这些犹豫有充分的理由。首先是2005年，《纽约时报》曝光了一个经美国政府批准，出于国家安全的考虑对AT&T的CDRs数据进行挖掘的项目，然后是近几年揭露的美国国家安全局PRiSM数据挖掘项目。于是，从2005年开始，这些数据的敏感性进入了公众视野。由此引发了一些诉讼，一部分案件中政府被认定为非法监视，另外一些则由于技术细节被推翻。

尽管政府无证搜索特定客户的CDRs数据与研究人员通过CDRs数据对人们的群体行为进行建模不同，移动运营商仍然对广泛公开这些数据很谨慎。即便大的数据集都是匿名的，并且诸如手机识别码和手机号码的客户识别信息被剥离，个人的识别信息仍然有可能被重新获得。以往研究显示，即便很多敏感信息如名字、地址和社会保险号码不可获取，通过另一些识别信息，例如性别、生日、邮政编码，一个人还是可能被识别出来。通过这三个识别信息与人口普查数据交叉分析，63%~87%的美国人口（取决于人口普查年份）可以被分别识别出来。

在一个对CDRs数据记录去匿名化的案例中，Sprint公司的研究人员使用的数据集覆盖了美国超过50个州的2 500万手机用户所拨打的300亿通电话。研究人员推断每个用户的常用地点，并将其与公开的普查数据相关联。研究人员得到的结论是，只要有足够的时间通过手机位置来推断常用地点，例如“家”和“单位”，基本上所有的用户都能被识别出来。他们建议，为了保证数据的匿名性，至少数据的时间和空间信息必须模糊。这意味着数据最好是在一天之内被采集，而不是经过一个月，并且不宜采集超过一个通信基站所覆盖地理范围的数据。

AT&T的研究人员试图让这些使得CDRs数据变得有价值的信息可以被大众获取，但同时注重保护隐私。普林斯顿大学的希伯伦·易萨曼和AT&T的拉蒙·卡塞雷斯以及他们的同事合作的一个项目叫作WHeRe（“从区域中找出单位和家庭”的缩写），致力于为特定城市开发手机用户的合成模型。通过使用AT&T实证数据中提取的空间输入记录和时间可能性分布数据，WHeRe可以为综合手机用户们创建数据。合成CDRs数据被应用在纽约和洛杉矶大都市区，并且通过了这两个城市中数十万手机所产生的数十亿的匿名位置样本的验证。这些合成的CDRs数据的主要优点是，它们维护了个人的隐私，因为没有一个数据是从真正的用户而来。另外，研究人员宣称WHeRe比其他手机数据建模的方法更精确。（例如，每天活动的范围落在真实范围的一英里范围内。）尽管在这个研究阶段仍然有很多需要做的，但这类项目会让人们获得更多的CDRs数据类型。

谷歌、脸谱网、推特

对这三个网络巨头不需要任何介绍，但本小节要提到三者采集的一些特定数据，以及总体来说三者之间获取这些数据的不同方式。谷歌的数据采集在早期就已经进行了扩张，那时他们仅仅监控用户的搜索词。因为它现在拥有并运营一系列不同种类的网络服务，包括Gmail电子邮件、谷歌+社交网络、Youtube视频网站以及Chrome浏览器，相比以前，它获取了更多个人用户的数据。如果Chrome浏览器的“即时”功能被打开，谷歌公司就会记录用户在搜索栏输入的搜索词和网址。另外，出于广告目的，谷歌还记录用户所观看的Youtube视频，在Google+上的操作，以及通过Gmail发送邮件的文本内容。

尽管普通人不能挖掘全部的谷歌搜索词和用户数据，但有一些方法可以获取到某些类别的数据。例如，通过加入谷歌的AdSense广告服

务，人们可以深入了解某些通过Gmail发送或接收的电子邮件中可能出现的特定的词或短语。个人可以通过互联网服务商进行查阅，获取在Gmail中收到这些广告的人数以及他们的位置信息，借此在使用这些词或短语的广告竞价中胜出。通过这种方法，人们可以了解特定的电子邮件流行语，并有可能发现不同国家的趋势或流行观点。

与谷歌相似，脸谱网也提供广告网络。当你发布一个广告的时候，你选择想要关联的关键词，这则广告会随之出现在个人信息中包含这些关键词的人的页面上。因为锁定广泛的人口统计资料或获取个人资料上的个别信息，如地址（市、州、省或国家）、人口信息（年龄层、性别、语言、感情状况）、兴趣爱好、教育和工作状况等，都是有可能的。尽管脸谱网并没有向广告发布者提供浏览广告的脸谱网用户的个人信息，但广告商会获得与广告条件相匹配的用户数量量级。

AdSense和脸谱网广告都有两种收费方式：通过每次点击付费以及每千次浏览付费。两者都可以通过各自公司的网页竞价购买。这两家公司提供广告的衡量数据，对试图卖产品或服务的人来说很有用。这些衡量数据包括广告被点击的次数，广告在不考虑点击时被展示给用户的次数（浏览），广告获取的点击量除以广告展示量（点击率），以及每获得一次点击需要付多少钱（平均点击付费）。两家公司同样也都提供总体统计数据，比如一个广告出现在电子邮件或个人页面的总数。这个信息可以用于判断不同人口对于某个产品的大体观点。

脸谱网同样提供一个应用程序编程接口，以便应用开发者在开发应用程序的时候调用脸谱网的用户信息。从这些应用程序中，应用开发者可以获取一个脸谱网用户在隐私设置中允许显示的所有信息，包括手机号码、联系人清单、状态更新以及其他个人识别信息。

一些最受欢迎的脸谱网应用，比如由社交游戏公司星佳开发的那些，不仅仅包含个人信息识别，同时还采集玩家的行为和习惯。数以

百计的手机应用拥有多达百万的用户，这在很大程度上是由于它们提供的某个特定功能会病毒式的传播，或者是可以提供特别令人满意的结果。简单地设计一款采集信息的应用远远不够，它必须变得足够受欢迎，才能提供足够样本数量以满足使用。然而，手机应用很难获得显著规模的样本群。

推特跟谷歌以及脸谱网都有很大的不同，它的大多数用户数据是公开的。推特提供一个名为“Twitter fire hose”的应用程序编程接口，可以获取连续的数据流。由于部分推文㊟包含位置信息，这些数据流可以用于发掘一个国家的民众情绪。第八章中更深入地介绍了专门分析推特数据的应用程序，以及如何利用它们发现国家层面上的趋势。然而，推特数据的分析所面临的最大挑战是如何从巨大的数据流中找到有用的内容。推文包含大量无意义的内容，比如对话的片段、链接、混杂的标签以及缩写等，这些对于分析来说都是挑战。

银行交易

银行交易是另一个国家层面的信息来源，但需要注意的是商业交易的数据基本上不可能被获取。银行常常使用数据分析来预防诈骗，或预测某个人更换银行的可能性，或者根据个人消费习惯和风险承受能力调整利率。对于研究人员和企业来说，客户的花销数据能够为迁移率以及策略制定的研究提供深刻见解。因为位置是与购物这个特定的行为结合在一起的，所以研究人员可以借此对一个人的经济行为有更全面的了解，甚至提供一些CDRs数据无法提供的信息。

2008年美国银行和麻省理工学院合作，让研究人员调查前者包含800万条客户交易的数据库。凯瑟琳·克鲁姆关注的是包含一万名客户在2006~2009年交易的子样本，主要指标包括交易数据、总量，是否使

用支票、借记卡或贷记卡，商户，商户类别码，以及交易是在线交易还是线下交易等。研究中涉及的交易总额达到每月300亿~350亿美元。

这类数据不是那些有求知欲的研究人员在实验室中每天都能查阅的数据。但像北卡罗来纳州的Tresata公司这类目标明确的初创企业，有可能可以成为大银行的合作伙伴，提供专门针对金融业的结构化和非结构化数据进行数据分析的平台。

Intuit公司旗下的金融公司Mint在金融交易数据流中也占有一席之地，这家公司提供的客户产品能直接将客户连接到线上消费追踪系统。Mint拥有超过400万用户，并能将个人消费模式可视化，以及提供信用卡等产品的促销信息。同时它还可以直接获取自家百万级别客户的交易信息。通过这些信息，该公司在推广产品时就可以根据个人的消费模式有的放矢。而且通过这400多万用户的交易数据，该公司可以对经济的健康程度进行综合判断；并且除了对某个客户信用卡使用兴趣这类简单推断之外，在更大层面上进行一些预测。

尽管想要获取公开发布数据以外的数据并不容易，但仍然有机会。虽然对大多数人来说通话数据记录都无法获得，但因为越来越多的手机运营商看到了与初创企业进行合作的价值——后者可以为运营商的记录提供深刻见解，就像有些大学也已经开始进行的类似合作；因此通话记录目前已经可以在有限范围内被共享。像WHERE这样可以生成合成数据集的项目，如果能够超越城市界限被广泛应用的话，也将成为一个重要工具。

谷歌和脸谱网向有兴趣打造广告并愿意为广告付费的公司或个人提供他们所需的数据。这些广告基本上都是面向大范围跨区域中使用特定搜索词或拥有特定兴趣的人。跨人口、跨区域的高度具体化数据，目前还是一个相对欠开发的资源。同样的，推特已提供更多有价值的现有信息，虽然这些数据的分析挑战更大。

大银行数据可能是这个层面上最难获取的数据，然而，移动运营商模型有可能也可以应用于金融业，如果该行业和那些可以提供某些价值的研究人员及初创企业合作。金融机构可能会发现一些现实挖掘从业者能提供有用的合成金融数据集。尽管大金融数据目前仍然不适合用于范围很广的应用程序，但可以肯定的是，这些数据绝对有可利用的空间。同样的，当这个层面上越来越多的数据被应用，数据卫士们可能会进一步开放数据的获取途径。下一章将会讨论这类数据的一些可能性应用的细节。

---

1. 更多信息见http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml。
2. 更多关于世界银行发展数据中心的信息见http://data.worldbank.org/about/development–data–group。
3. 推文：指在推特上发的内容，包括文字、图片等。——编者注

第八章
让大数据发挥作用的最佳方法

国家数据库对于理解怎样进行资源分配和设计政策非常重要。但更重要的是要找到让数据发挥作用的最佳方法。人口普查、data.gov网站以及世界银行的数据都提供了关于国家规模人口的重要解读，但这些观点往往是静态的。这些资源由于受到传统的数据采集技术的限制，提供的只是数据在一定时间和地点条件下的简单描述。通过数据的可视化，如时间线、地图、图表等，政策制定者以及非政府组织可以看到这些大量的、动态的数据如何通过不同的方法被划分、切割、交叉引用。数据可视化的结果可以被用于为贫穷和饥饿的人群提供更好的服务，更好地理解选民行为，以及对之前被掩盖的行业进行早期投资。

通过可视化来发现这些潜在的应用有很多方式。谷歌提供了一个名为"谷歌公共数据管理器"的工具，使用世界银行以及其他数据集，让人们对一段时间内不同的统计数据进行比较。Data.gov网站链接了很多美国政府和非政府组织的可视化数据，并常将这些数据在地图上进行叠加。

一个名为Development Seed的数据可视化组织，相较于其他数据来源选择了使用世界银行的数据，并创建了一个互动地图。而它的子计划Mapbox，则让其他人可以通过开放数据简单快速地制作互动地图。Development Seed组织进行的数据可视化通常有特定的目标，如"识别关键问题、用战略思维发掘机会、建立解决方案及设计交流"。例如，他们的可视化项目关注非洲之角饥饿问题的发展趋势，以及阿富汗的公平选举。

在国家层面，数据集之间的一个主要不同点是它们的更新速度。人口普查数据和世界银行的数据更新得很慢，然而，手机数据的更新则非常快。第七章提到的CDRs数据能够近乎实时地提供人们的移动性信息。将这些新的移动性信息和手机CDRs的历史数据结合起来，工程师可以获得人口移动和活动网络的总体情况。通过这些信息，类似将贫民窟的人口流动和主要粮食的价格相关联的一类研究就成为可能。这个模型最终可用于辅助预测人口的大量涌入，从而帮助管理者在正确时间正确地点分配正确的资源。

本章除了探索人口普查数据和CDRs数据的综合运用之外，还介绍了使用谷歌和脸谱网的广告网络来实时发掘国家层面多数派观点的方法。从某种角度来说，这种方法像是建立了一个个人的民意调研——提出一个特别的问题，并使其含有能够使人想起你的广告的搜索词或关键字，然后只需要简单地监控相关指标即可。通过这种方法，人们可以通过创建一系列的单变量民意测验来获得某个特定人群的观点和感受。比如说，人们可能会跟踪那些在某个特定的政治候选人脸谱网页面上点“赞”的人。

另外，本章还介绍了联合国的全球脉动项目，该项目通过采集推特公司的全球第四大市场印度尼西亚的推文信息，来更好地了解当地居民承受的压力。通过推文内容分析，这个项目可以分析人们有关燃料、食物、债务以及其他能暗示压力的因素，从而预测危机。

最后，我们讨论了新兴的金融数据分析领域的应用。相对难以获得的金融数据，特别是那些实时采集和分析的数据，可以展现一个国家即时的经济健康状态和福利状况。

人口快照

国家制定公共政策时，拥有不同区域的需求、资源以及人口组成等大量的数据是很重要的。

世界银行有200多个国家提供的超过7 000个指标，涵盖了从教育水平到军费开支等多个方面，从中获取最重要的信息可能会有一定的困难。为了弄明白这些数据集，世界银行提供了应用程序编程接口，让其他机构甚至个人软件开发者都可以接入这些数据。谷歌的公共数据管理器是一个方便接入和使用的例子，通过把世界银行的数据和其他60个公开数据集相结合，谷歌的这个互动工具可以让人们看到一段时间内的发展趋势，也可以比较不同的国家和不同的统计数据，还可以回答诸如“哪个国家的识字率正在提高？”“这会怎样影响GDP的改变？”之类的问题。

华盛顿市的一个数据可视化组织Development Seed采用的方法和谷歌不同，它们关注某些特定的问题，并创建风格化、视觉上引人入胜的地图和图表。另外，Development Seed网站的开发设计人员与合作机构紧密互动，力求寻找以可视化回答问题的最佳途径，并最终影响到政策制定。该组织最近的产品是Mapbox应用程序，该程序是研究者为了使不同的开放数据能够更容易、更快捷地被公众获取所做的一项尝试。

Mapbox项目的一个很有分量的案例就是对非洲之角2011年大饥荒的影响进行的制图，这次大饥荒是由干旱、冲突以及食物价格上涨等多方因素共同造成的，它影响了1 300万人口。它使用了来自美国国际开发署和联合国人道主义事务协调厅的数据，该组织创建了当前以及预计的饥荒程度和干旱状况的地图。这些地图可以被修改，这样人们就可以在已有的内容上增加不同的信息图层。但其主要目的是让更多的人了解这次饥荒，并号召人们向世界粮食计划署为这次特定的需求捐钱。

记者和政策制定者一样，在众多指标中，都对能够反映社会进步或者止步不前，人口的总体变化，经济，政治面貌，宗教信仰等的发展趋势以及历史数据感兴趣。Development Seed组织和美国公共广播电台合作，通过聚焦西班牙裔的人口增长以及他们选择在哪些地区生活，来展示美国的人口变化情况。这类数据在政治、社会或经济事件的报道中很重要，它们可以提供观点支撑以及事件背景。

Development Seed组织的另一个项目是基于美国的机会指数创建的互动地图，它展示了各州、郡在不同指标下的不同表现，这些指标包括失业率、家庭收入、高速网络订购水平、受教育水平、志愿服务水平以及获取健康食物的便利度。这些指标最终可以用于帮助人们选择工作迁居地，并鼓励政府或企业在不同部门推行改进计划。

Development Seed组织还创建了一个气候变化地图，展示了未来的全球温度和降雨量变化情况。当将这些预测和世界银行的产业类型以及特定区域的人均收入数据结合起来时，这个地图就可以被用来预测气候变化相关的经济问题，包括类似洪水这样的自然灾害，以及导致作物歉收这样的农业气象灾害等。

当今世界，对于寻求减轻气候变化带来的严重后果的活动家、科学家以及政治家来说，当地图和其他与气候变化相关的数据来源结合在一起的时候，可以成为一个更有力的工具。而在一些气候变化会带来较大影响的地区，有些投资人看到了新的商机，也可以利用这些数据。比如，对于建造堤坝和其他防洪设施的建设公司来说，海平面上升就是个好消息。然而，总体上来说，这些地图和数据集具有打破国界的潜力，可以被用来促进降低温室气体的排放和开发对气候影响较小的技术。

人口流动数据的重要意义

尽管人口普查数据、世界银行以及其他政府数据集，对揭示缓慢变化的社会和经济趋势很重要，但这些数据缺乏粒度和时效性。CDRs数据可以填补这些空缺。因为CDRs数据能够提供粗略的移动状态和一整个国家和地区的社交网络，它们可以用于人类行为的建模，从而快速识别行为变化，并且预测未来的行为改变。使用CDRs数据的基础研究可以识别人们在现实生活中，与其他人建立联系的模式以及他们出行的方式，最终，CDRs数据的相关分析可以应用在特定的场景分析中。例如，CDRs数据可以帮助政府部门更好地了解贫民窟等临时居住地的动态；以促进更好地分配资源；还可以用于预测诸如地震之类的自然灾害会对这些居住地产生什么样的影响。

2007年和2008年，艾伯特·拉斯洛·巴拉巴斯和他的同事在《自然》杂志上发表的两篇被广泛引用的论文，就是关于CDRs数据比较有影响力的早期研究。这两篇论文阐释了CDRs数据可用于人类社交网络以及移动性动态研究的潜力。

2007年的项目使用了未公开服务商提供的超过40亿用户的匿名通话记录，来研究社交网络的进化。通过这些数据，研究人员研究了小规模人群（朋友圈、家庭、专业圈子）和大规模人群（比如学校、机构和公司这类人口组成相互重叠的社区）一段时间内的不同动态。如果在给定的时间段中，已有成员被替代得越少，那么小规模群体就越稳定；而大规模群体反而在经历规模和构成波动的情况下更加稳定。

2008年的研究关注的是10万个匿名手机用户的移动动态。以前，人类的移动性预测一般用列维飞行（Levy fight）[注] 和随机游走模型来假设一个人活动的时空规律。从CDRs中采集的位置数据显示，尽管不同的人过去的出行模式不同，从个人的角度来说，人们倾向于追寻简单、可重复的移动模式。2010年，巴拉巴斯和他的同事在《科学》杂志刊登了一篇论文表明，即便出行的频率、出行目的地不同，93%

的人的出行是可以被预测的。而且即使是那种出行最没有规律的人，研究者通过CDRs数据模型，至少可以预测其80%的出行。

巴拉巴斯及其同事的研究揭示了CDRs数据集的巨大潜能。当应用到真实生活中的问题时，CDRs数据可以解释清楚用其他方式很难调研的社会、环境，或者经济状况。

全世界有超过10亿人生活在20万个贫民窟、人口密集且肮脏的贫困区域、破败失修的临时住所，以及其他非正式的居住环境中。这些居民中使用手机的有数亿人。因此，可以通过手机数据来掌握这些居住在贫民窟中的人每日、每月和每年的移动模式。这些移动模型可以帮助确认并改善现有的贫民窟动态理论。比如，贫民窟中的人口流动会怎样影响附近的城市？是什么让贫民窟中的人口增加或是减少？通过这些问题的答案，不同机构可以分别找到帮助人们的方法，或者设计有效的诱导机制帮助人们找到更长久的住所。

埃米·维索洛夫斯基和内森·伊格尔通过手机记录研究了位于内罗毕市中心西南的基贝拉贫民窟，这也是世界上最大的贫民窟之一。研究人员发现了人们进入或迁出基贝拉的迁移模式，这让他们可以推断其居民的工作的地点和部落的从属关系。一个惊人的发现是：贫民窟中的居民的迁移率很高，接近一半的居民每个月换一个夜间居住地点。当试图推断居民的新居所的位置、部落关系、工作区域时，研究人员发现每个月的此类信息很少有重合的时候。研究者将众人在白天聚集的一些地方——很可能是他们的工作地点，称为“经济跳板”，并且长期监测这些地点。

研究人员认为，通过识别这些经济聚集点的模式，可以建立贫民窟动态的预测模型。比如，可以预测新贫民窟的位置，或是一个贫民窟是怎样发展起来的，以及贫民窟对周边城市区域的发展会产生怎样的影响。当地政府可以分配给贫民窟居民用于保障其福利的，只有少量的宝贵资源。对贫民窟将会如何随时间变迁有更好的了解，有助于

城市管理者确定固定基础设施投资的最佳位置，比如确定哪里适合安装下一条净水管道，或是哪里适合设置一个新的公共厕所。

除了提供针对像贫民窟这样长期问题的移动模型，CDRs数据还可以用来研究如何采用最佳方式来应对像地震这样的严重灾难。伊格尔和他的同事分析了2008年2月在刚果民主共和国基伍湖区域发生地震时通话记录中的变化。在此研究中，一个基线活动的CDRs数据变化被用于确定地震发生的时间以及震中所在地，并找出反常通话行为持续存在的区域，这些区域可能正是需要援助的区域。CDRs数据可以用于灾害分析表明自然灾害的预测性分析可能可以依靠现有的电信基础设施来进行，这使人们制订缓解灾害带来的饥荒和疾病传播的事前预案成为可能。

2012年，在撒哈拉沙漠以南的非洲大陆，疟疾导致了超过100万名5岁以下的儿童死亡。当地政府和健康组织没有成功抑制该疾病的一个主要原因，是没有深入了解疟疾感染者是如何在区域内医疗不足的非正式社区中迁移的。在一次史无前例的数据分享合作中，肯尼亚卫生部、当地移动运营商以及一些世界最知名的疟疾专家都参与其中，伊格尔和他的合作者得到了令人信服的证据，证实了在定量分析疟原虫在区域寄主人群中传播时，手机数据是最准确的方法。

现在，人们已经可以衡量国家层面上人的迁移率、规模和方向。手机数据结合详细的寄生虫风险地图，让研究人员有史以来首次可以确定一个给定区域招致疟疾的风险，从根本上确定疟疾发生的热点区域。发现疟疾热点改变了疟疾防控资源在整个肯尼亚的分配方式，政府部门现在按照新的战略部署大量根除疟疾的资源，这些新战略相较于传统战略，包含了对人们的真实移动模式的数据分析结果。研究人员将传染性疾病包括疟疾的专业知识与手机数据相结合，创建了一个成功追踪和预测疾病传播的新方式，使得资源分配更智慧、更有效。

让广告更聪明

当多数人想通过谷歌和脸谱网做宣传时，他们希望的是为自己的公司带来更多生意。但在推动销售外，谷歌和脸谱网的广告还为巨大的数据储量打开了一扇窗——提供网络上实时的观点和意见快照。谷歌的AdSense产品和脸谱网广告都提供了某广告的投放用户衡量数据。因为这些广告跟选定的谷歌关键词相关联，也可以与脸谱网的基本信息和兴趣爱好相关联，所以创建这些广告或活动的人可以获取收到广告的人们的（匿名）数据。

假设这样一种场景：你创建了一个AdSense的活动，其关键词为“流感”和“生病”。然后随着时间的推移，你可以看到关于这些关键词的指标是如何变化的，获得这些数据的途径包括搜索或查看Gmail中的邮件。比如，当秋冬季节流感在北半球开始暴发的时候，看到你的广告的人数会逐渐上升。通过查找网络供应商数据，你可以绘制流感症状在一个国家内的传播的地图，找到热点和传播模式，甚至可以预测下次暴发的规模、地点或者其他动态。

这种方式的流感追踪已经有一些先例。风瑟·埃森巴赫在2004~2005年，于加拿大的流感季节使用AdSense追踪流感。他创建了一个当人们搜索“流感”或“流感症状”的时候会出现的广告。该广告的内容是：“你感冒了吗？是否发烧、胸闷、虚弱、疼痛、头痛、咳嗽？”，并链接至一个普通的医学教育网站。他发现这个方法“比起传统的定点医师发现类似流感疾病并上报的方法更加及时准确，且便宜得多——整个流感季节仅花费了365.64加币。”

除了可以追踪一个国家的流感传播情况（本书的第五部分还介绍了全球疾病追踪器，比如谷歌流感趋势）,AdSense的方法还可以用于政治分析或其他需要民意测验的领域。2012年的选举中，一个广告

在“我希望奥巴马赢”作为关键词被输入时会出现，另一个广告在“我希望奥巴马输”被输入时会出现，它们就可以作为采集两种观点的渠道。

在本书写作时，脸谱网的广告系统和谷歌还有所不同，广告投放者从脸谱网获得的是用户个人信息的一些特定部分。根据脸谱网广告的协议条款，广告商可以把用户个人信息中的特定内容作为目标，包括地理位置（城市、州、省或国家）、基本信息（年龄范围、性别、语言、婚姻状态）、爱好和兴趣、受教育程度以及职业。在这种情况下，广告商并不能获得用户的状态更新或评论中的词汇或短语。另外，他们也不能获取照片的相关信息，比如位置信息、活动内容等。（需要注意的是，条款会不断地修改。脸谱网的移动应用强调了图片的分享功能，表明了脸谱网对于获取移动设备拍摄的照片有很大兴趣，这些照片附带了丰富的元数据。）

尽管脸谱网对广告植入有很多限制，但通过它，人们可以有效地获取不同人群中的特定观点。例如，在2012年选举期间，你可以设定一则广告，让它显示在标记了“喜欢米特·罗姆尼”或在兴趣中提到他的名字并且受过高等教育的用户的页面上。

到目前为止，对于匿名挖掘在线广告的衡量数据以获得接近实时的调查对象的观点这种想法，学术界尚未有广泛的探索。但这个方法被证明可以让个人有效、及时、便宜地获取流行观点和其他隐藏在谷歌和脸谱网的独占数据背后的要素。

通过推文识别危机

与挖掘脸谱网和谷歌的数据山所需的曲折方法不同，推特的开放生态系统让数据挖掘更加直接：采集推文，过滤并分析有关压力、危机或其他观点的指标。2011年10月，联合国秘书长办公厅倡议的“全球

脉动”项目，出版了对印度尼西亚和美国的推文进行研究的结果。其目标是更好地理解公民的观点，并为现有的政策研究增加筹码。

该项目把推文进行分类，并对公民关心的问题进行了量化分析。研究人员关注的主要衡量指标包括关于某个特定话题的推文的反常增加或减少，比如停电；推特对话主题每周或每月的变化趋势；某个特定话题在一段时间内的容量特征；不同子话题的比例，比如非正式和正式借款；以及推特聊天的方式和其他不同来源的数据比较。

“全球脉动”项目的合作方Crimson Hexagon，是一家采集和分析公开社会媒体数据的公司。根据这家公司的分析，印度尼西亚是推特公司的全球第四大市场，其国民每天发布55亿条带有位置标签的推文。在“全球脉动”项目中，Crimson Hexagon公司获取了2010年7月到2011年10月所有公开发布的推文，通过使用哈佛大学定量社会科学研究所开发的观点分析运算法则，来分析文本中的关键词以及这些词之间的关系，研究人员可以了解引起话题来源的主题类型和相对强度，如印度尼西亚这段时间内的债务、食物价格和供应量以及燃料等问题。印度尼西亚的一些推文话题被用来与美国的推特对话相比较，借此观察两者差异。

其中的一个发现表明两个国家有关燃料的推文有所不同。美国关于燃料的推文主要关心的是价格，而印度尼西亚的推文关注的则是是否有不同类型的燃料，比如汽油、柴油、石油和煤油。这个区别意味着，进一步的研究需要区分燃料话题的相关子话题。另外，研究人员发现印度尼西亚关于粮食价格的推文的数量与通货膨胀统计数据密切相关。

研究人员得出的结论表明，数据和文本分析不能作为衡量一个人的长期目标和关心的问题的标准。但他们发现，推特数据作为筛选观点的一个总体方法非常有用，特别是用于研究一个国家公民当前所关心的问题。

挖掘金融的未来

虽然对于一个研究人员或企业家来说，最具有挑战的可能是去获取匿名金融数据的大数据库，但这些数据可以提供关于一个国家的经济状况信息，甚至还可以预测未来的经济危机。当然，大银行拥有并使用这些数据来判断客户的周转率，决定哪些人可以贷款、哪些人具有良好信用以及使用什么样的利率。

但如果这类数据可以更广泛地供大银行以外的研究人员或企业使用，对于所有规模的行为经济都将是有利的。麻省理工学院的凯瑟琳·克鲁姆（第七章也讨论过）的工作可以让我们了解一些这种可能性。2008年，克鲁姆获得权限接入美国银行的一个匿名数据库，其中含有8 000万客户的金融数据。她的分析主要是关于这些客户的购买模式。克鲁姆发现，经常把钱花在杂货店和餐馆的人们的行为最容易预测，而那些把钱花在加油站和快餐连锁店的人不易预测。她还发现在一次购物出行中，富有的人倾向于在不同的商店消费，而那些不那么富有的人，在一次购买行为中基本上只去一家商店。

这类分析只是冰山一角。当银行数据和其他类型的数据结合起来，比如地理数据和人口普查数据，人们可以在一个空前规模上对国家的经济版图进行描绘。比如，人们可以预测工作岗位的增长和下降，或者将消费模式与消费信心指数连接在一起。后者可以提供最新的即时消费信心指数。这些数据可以用于预测哪些产业会受到经济衰退的影响，还可能预测它们受影响的程度。大数据驱动的指标可以帮助企业和政府做出可能缓和将发生的经济危机的决定。

个人金融初创企业Mint创办于2006年，已拥有数百万的最新金融交易信息。该公司为用户提供简单易用的界面，让用户可以通过手机应用程序或电脑网页浏览自己所有的金融账户信息。除了向用户提供他们的基本金融信息，Mint还提供理财建议，比如说，根据用户需求

推荐信用卡或储蓄账户。2009年，曾开发了广受欢迎的个人金融软件Quicken的Intuit公司收购Mint的时候，也同时购买了所有Mint的客户数据。

现在，Mint的主要业务是向客户提供他们的消费和储蓄信息，并推断他们未来的习惯以便向客户推荐信用卡、健康保险、旅行等的最佳促销活动。2010年，Mint公司启用了一个可视化工具——MintData，它可以让人们查看所有Mint用户的匿名总消费信息。例如，人们可以看到大家在俄克拉何马州的餐馆与在华盛顿的餐馆平均消费水平的比较。MintData项目已经被终止了，但在2013年本书写作期间，Mint公司已经计划开发一个相似的可视化工具。

如果Mint的百万客户的匿名金融数据都公开的话，那么除了一部分Intuit公司的研究人员和与该公司合作的经济学家之外，会有更多的人可以对这些数据进行研究。发现金融数据中的趋势，可以帮助预测经济危机，也可以帮助建立基于数据的改善经济指南。将金融数据和其他数据集配对起来，甚至可以通过回答一些简单问题，比如在哪个区域消费者购买了更多的橙汁以预防感冒，来预测流感的传染。虽然这些都还是金融数据分析的初步应用，但拥有获取这种类型的数据对人们来说已经很令人激动了。

国家层面的数据有很多种使用方式，其中的一些对政策具有很重要的意义。数据可视化提供了一种令人信服的直观方式来讲述数据背后的故事。当通过这些可视化的图表、动画或是互动数据工具可以很容易地发现数据的变化趋势时，政策制定者们就不容易忽视可能发生的危机。数据可视化或其他信息分析工具总是希望能获得最新的数据，而这正是CDRs数据可以发挥作用之处。除了提供高密度的大众移动信息，CDRs数据还可以提供几乎实时的信息，从而帮助相关人员尽可能快速和有效地处理自然灾害等危机。

为了获得对国家层面的感知，你可以从我们这个时代的网络巨头公司获得该层面的数据。你可以通过接入谷歌和脸谱网的广告服务网络，发布广告并分析挖掘其结果指标数据。推特具有大量的公共可获得数据，同样也可提供丰富的大众观点。通过语句文本解析算法，可以从推特每天亿万条140个字符以内的海量内容中抽取多样化的信息。然而，所有这些方式都是为了确保你提出了正确的问题并验证你的发现。第九章将涉及一种重要的双校验机制：将个体重新放回更大范围的全球链中。

---

1. 列维飞行，根据法国数学家保罗·列维命名，是一种随机游走模型。——译者注

REALITY
MINING

Using Big Data
to Engineer a Better World

第五部分
世界数据

很多趋势可以通过对搜索词的流行程度、频率和周期性等进行调查来获得。谷歌提供了一个名为“谷歌公共数据管理器”的工具，使用世界银行以及其他数据集，让人们对一段时间内不同的统计数据进行比较。Data.gov网站链接了很多美国政府和非政府组织的可视化数据，并常将这些数据在地图上进行叠加。

REALITY

MINING

Using Big Data

to engineer a Better World

第九章
大数据能为世界带来什么?

现实挖掘影响深远的应用之一是预测疾病和追踪传染病。在当今的全球化世界，致命疾病会灾难性的传播至空前多的人口。本书的最后这个部分介绍了全球层面的数据，并着眼于更好地了解疾病是如何在我们这个大规模连接的世界中传播的。由于数据集可以告诉我们人们是如何迁移、他们在网络上搜索什么以及他们的感受，因此我们有机会创造了一个基于信息的概念来说明世界是如何运转的。

正如世界各国共同参与协调人口研究一样，世界银行、联合国以及其他国际机构已经启动并持续采集跨国数据。“千年发展目标”提出了一系列到2015年应实现的8个目标，包括使极端贫穷和饥饿人口减半、控制艾滋病毒/艾滋病的传播等。在这些目标的激励下，这些国际

组织建立了国家内以及国家间的标准化协调采集数据计划和指导方针，并启动了国际数据采集行动。

当然，正如在第四部分中讨论的关于国家数据普查一样，全球数据普查也是静态的。因此，在对疾病扩散进行建模时，一个重要的步骤是量化人口的迁移；同时，我们还必须关注网络数据集，比如航空旅行和远洋航线。每年有上亿乘客乘坐飞机，每个乘客都有可能把疾病带到新的人群中，而且尽管海运并不会有航空乘客那么多，但海运的食品和货物在不知不觉中也可能含有入侵物种，比如有问题的病毒、细菌或寄生虫。

在全球层面上，最有吸引力的数据来源是互联网搜索查询，它为研究人类行为和环境情况提供了一扇窗户。谷歌的匿名搜索词数据库是该层面最强大的数据来源之一（一个复杂的事实和需要关注的问题是政府是否会非法获取这些数据），该数据库在用户没有登录自己的谷歌账户的状态下不会识别特定的个人，而是通过互联网供应商的地址来识别该用户。谷歌自己在一些有用的工具中使用并公布了部分数据，比如谷歌流感追踪系统和谷歌搜索解析。这些工具表明，任何人都可以将一段时间内的搜索词的受欢迎程度可视化，并下载有用的数据以便进一步研究。

推特也可以用于全球层面。本章中我们讨论了一些接入推特的fire hose数据接口的不同方法，该数据接口每天产生成千上万的推文。

但是，通过这些全球数据，有可能会导致草率判断现实状况。在大规模数据上建立的模型分析，应该尽可能经常通过个体经验来证实，为自上而下的判断提供自下而上的经验检验。为此，国际化的初创企业Jana开发了一个手机装置来采集广泛观点。该公司在全球超过100个国家中，为手机用户提供赚取少量通话时间的机会。一种常用的方法是让人们通过完成手机问卷来赚取通话时间，这种问卷可以包括非常广泛的话题，简单的问题比如某种特定的产品在当地商店中是否

有售？答题者最近是否有类似流感的症状？这类自陈式调查数据提供了一个重要的、非侵略性的方法，可以用来验证由人口层面数据而产生的推断分析。

本章简单介绍了全球层面数据采集的不同方法，并着眼于大数据最吸引人的应用之一：追踪、预测乃至消除流行疾病。

全球人口普查

像世界银行、联合国统计委员会、经济合作与发展组织以及国际货币基金组织这样的国际组织已经积极地确保所采集的关于参与国的数据达到一定的标准。据此，它们共同合作为统计工作开发合适的框架以及最佳实践政策，并在追踪记录数据交换过程和方法的指标上达成了共识。世界银行在这些最佳实践的约束下，根据来自不同国家统计系统的数据来汇编国际数据集。另外，它还资助在全球采集数据的项目，比如MAPS计划（马拉喀什数据统计行动计划）以及PARIS21联盟（21世纪发展统计伙伴联盟）。

MAPS计划共包括6个行动，旨在改善国际和国家统计系统。在“千年发展目标”的驱动下，MAPS计划包括三个专门关注全球层面数据采集的行动：一是为所有低收入国家规划统计系统并规划国家统计发展战略，直到2006年；二是保证所有发展中国家参与2010年的人口普查；三是设立国际家庭调查网络，用于采集以家庭为单位的全球社会经济数据。

PARIS21联盟则是一个决策者和分析师之间的国际合作。该论坛成立于1999年，促进、影响并帮助了很多国家通过最有效和最有用的方法使用统计数据。它帮助“低收入和中低收入国家设计、执行并监管国家层面的统计发展战略”——其本质上就是建立统一数据采集标准的

一种方法，“并让国家拥有和生产所有千年发展目标所要求的指标数据”。

得益于世界范围内不同组织的协同努力，跨国数据和全球人口普查数据在符合合理预期标准和连贯性的情况下，现在已经可以被获取。特别是那些可以通过世界银行、经济合作与发展组织以及国际货币基金组织的网站在线获取的数据，可以让人对全球整体经济状况有个准确而快速地了解。

航运和航海的足迹

也许没有比国内和国际航空航线，更能够在全球层面上代表人类流动的途径了。2011年的美国，共有7.3亿乘客乘坐商用飞机出行。同年，全球飞机共运输了28亿乘客。人们从A地点到B地点的路线选择可以和人口普查数据、疾病地图以及其他数据集交叉验证，用于了解信息、交易和疾病的传播。

这些流动路径数据的一个主要来源是国际航空运输协会，它们运营的数据库采集了2000年以来的国内和国际航线的信息。该数据库的信息来自130多家航空公司，可以覆盖全球每月约90%的飞行计划。这些数据可以在国际航空运输协会的网站上下载，订阅12个月的数据大约要花费1 000美元。在2006年，美国国家科学院院刊的一篇论文中，维多利亚·科利扎和她的同事们使用了国际航空运输协会数据库中2002年的数据，包括由直达航班连接的成对机场以及任意指定航线的剩余座位。他们的研究成果形成了一个包含3 880个端点（机场）以及18 810条连线（在机场之间流动的乘客）的网络，研究人员利用这个网络建立了全球疾病传播模型。

另一个航空乘客数据的来源是UBM Aviation公司，它们出售有关全球航空业的数据、分析和咨询服务。截至2012年7月，UBM Aviation的数据公司OAG的数据库包括超过900条的航线以及超过4 000个的机场，该公司同时提供1979年以来的历史航班信息时间表。OAG记录的航线报告内容包括某次航班共有多少个座位，以及该航班的出发城市和目标城市。很多研究人员曾经使用OAG数据库来预估疾病传播。这些数据库可以在OAG的网站上购买，价格由不同的国家、数据类型以及订阅模式决定，比如在美国，“OAG全球飞行指南”一年的订阅价格为845美元。

除了航空以外，货船也提供了一些研究人类移动性的切入点。虽然海运运输的主要是货物而非乘客（货物运输占据世界交易的90%），但航运网络仍然在物种入侵和疾病传播中扮演着重要的角色。

2001年，船舶和港口开始使用自动识别系统（简称AIS）设备，从而自动将海上交通的来往航线电子化。尽管自动识别系统主要用于防止碰撞和提高港口的安全性，该系统的附带用途是可以用来建立一个庞大的航运网络数据库。海洋网络数据库是一个关于航海统计的在线记录，其内容包括船舶的到达和出发数据，订阅价格在630~1 4000美元不等（取决于度量标准以及用户人数）。巴勃罗·卡鲁扎以及他的同事通过这个数据库获得了2007年的历史自动识别系统数据，并借此追踪了超过16 000艘船舶以及近千个港口，连接了超过36 351对有直接航线的港口。基于2007年主要货船的行程，研究人员发现了在干散货运输、集装箱货运以及油罐运输之间出现的三种截然不同的运输模式。比如，集装箱货运遵循有规律的重复航线，而干散货运输和有关运输在港口之间的运输则相对较难被预测。

谷歌“趋势”

根据采集了谷歌官方历史数据和Comscore数据的Statistic Brain公司提供的信息，搜索巨头谷歌平均每天收到47亿次网页查询，仅2011年一年就有1.7万亿次搜索。这些搜索是了解人们的兴趣以及所处状况的线索。该公司甚至将流感暴发和搜索词频率联系起来（第十章中会进一步解释）。同样的，很多其他趋势可以通过对搜索词的流行程度、频率和周期性等进行调查来获得。

谷歌提供了一个名为“趋势”的免费公共工具，让任何搜索网页搜索的人都可以看到搜索随时间的变化。该工具使用来自上亿匿名用户的搜索查询数据，而且只要用户登录自己的谷歌账户，就可以下载。

社交网络的全球数据

针对全球层面在线数据的讨论是无法避开推特和脸谱网的。正如在第七章和第八章中提到的，推特拥有一个名为fire hose的数据接口，让研究人员和企业可以获取绝大多数的推文。已经有大量的书讨论如何让推特数据有意义。本书中，我们提供了一些在进行定制分析时很有用的数据来源选项。

2012年中期，推特宣称其全球超过1.4亿用户每天发布的推文超过了4亿条。2012年8月，推特发布了他们的认证合作项目，特别提到了DataSift、Gnip、Topsy等12个合作伙伴公司。每个公司都有不同的目标和方法，但都可以深入获取推特数据以及关于数据的深入了解。这些公司反过来用不同的方法分析数据，然后将预分类过的数据集和服务提供给其他公司或个人，这些公司或个人往往对通过推特表达的不同观点有兴趣。如果想要找到关于2012年总统选举的推文和对话，DataSift公司可以使用自然语言处理以及其他过滤条件来采集和整理这些推文，并提供每月3 000~15 000美元价格不等的产品，这取决于收到数据的单位和所提供的服务。

对于那些选择自行接入推特的人来说，可以通过推特自己的应用程序编程接口来进行现实挖掘。而据推特公司的说法，这些应用程序编程接口也在不断地更新。一个可用的应用程序编程接口是用来查询推特内容的，比如关键词、涉及特定用户的推文或是来自某个特定用户的推文。不过，这个应用程序编程接口有局限性。它不是一个推文的完整索引：它不能搜索一个星期以前的推文，而且搜索词在复杂性上受限。推特的流式应用程序编程接口比搜索应用程序编程接口好用，比如，允许追踪更多的关键词，搜集带有特定区域地理标记的推文。所有推文在一定的比例限制下的随机抽样片段，就是通过这个应用程序编程接口抽取出来的。

正如第七章中提到的，获取脸谱网数据的途径之一是广告分析，但还有另一种方法是通过该公司为开发人员设计的开放图表工具。2010年，该公司提供开放图谱协议，向大众提供了用户社交图谱中公开的部分，包括好友、照片、活动和页面。社交图谱中的每个对象都有唯一的ID，并且可以通过图谱应用程序编程接口调用ID。图谱应用程序编程接口的标准版本只允许用户在每次查询时检索一条信息，批处理工具让用户一次可以发起多达50次的查询。社交网络上的流行应用程序的开发者，在用户授权该应用获取数据的情况下，可以通过连续的批处理快速地检索数百万脸谱网用户数据。

脸谱网和推特都是正在发展中的公司，它们都在改变获取一定数量或类型数据的方法。关于全球层面上人们的思考、移动以及行为，这两家公司都向研究人员提供并且将持续提供新观点，因此它们也是现实挖掘所需数据的重要来源。

现实挖掘的实际核查

通过大量数据对世界进行自上而下的、上帝视角的观察，可以使人们获得前所未有的了解世界各个角落出现的趋势的途径。但如果从全世界千万亿字节级别的数据中得出的结论没有复核机制，那么从这些数据中建立的模型可能没法像人们希望的那么有价值。手机除了在这个层面上通过CDRs数据提供对于人们的移动性的观察（见第四部分），还可以作为个人状况和环境的有效核查工具。

关于这些现实核查的重要性的一个早期案例，是关于卢旺达的霍乱暴发的研究。2009年，内森·伊格尔在卢旺达与公共健康社区以及当地电信公司合作，试图证明从CDRs数据中得到的人们的活动情况能够预测疾病的暴发。研究人员相信，如果疾病将要暴发，一个主要的表现是在被感染的社区中的人口活动会减少。他们怀疑人口活动的减少是由类似流感的症状导致的，而且似乎是预示一周后霍乱暴发的最初信号。

然而，这些社区中活动的减少被证明并不是因为人们受到类似流感症状的困扰，而是由于暴雨冲毁了道路。模型发现活动减少是因为洪水而不是因为霍乱。尽管洪水是霍乱暴发的先兆，但活动的减少被证明与流感症状的困扰无关。但仅从全球的视角来看，人口流动性低与疾病之间的关联度比起与基础设施问题之间的联系要更为突出。

得益于伊格尔和全球移动运营商之间的良好关系，他在2009年成立了Jana公司（前身为txteagle），该公司让国际品牌可以方便地通过手机与人们打交道。到2013年中期，Jana公司在全球已有35亿订阅用户。Jana的主要目标之一是采集消费者对于新兴市场的观点，在这个领域传统的市场调研往往不够有用或高效。Jana公司提供激励措施以吸引人们加入研究小组，提供他们的观点和意见的用户可以换取额外的通话时间。该公司正在建立一个自助界面，让研究人员可以向超过100个国家的手机订阅者发送定制的调查问卷。通过这种方式，Jana提

供了一种直接从35亿手机订阅者的样本中获取观点的方法，为全球的大数据分析现实核查提供了入口。

有史以来，人类第一次拥有在全球层面看到自身行为的能力。不同的人口普查、出行和航行网络、基于网页的搜索、社交网络行为，以及手机的使用正在这种能力中扮演着重要的角色。出现这些变化和发展，一部分原因是人与人之间的现实连接和数字连接都大大地增多了，这也导致了疾病传播的速度变快和范围变大。下一章将会深入探讨关于流行病追踪、建模和预测领域中的一些可能的应用。

第十章
明天会更好

当今世界被数据覆盖，从飞行网络到通话数据记录以及网络搜索和脸谱网状态更新。在全球层面建立一个系统的方法来改善全球健康状况也许是大数据最有价值的一个应用。作为本书的最后一章，本章专门探讨使用全球数据来识别以及阻止传染病传播的方法，这些传染性疾病包括流行性感冒、疟疾等。通过前述章节中提到的数据，我们来看一下可以用它们建立特定疾病在全球传播模型的各种方法。

疾病通过人、昆虫以及其他载体传播。为了了解疾病是如何传播的，将载体的移动状况进行定量很重要。一种采集移动状态信息的方法是使用航空和海运路线上的数据。在上一章中已经介绍过，这种粗略地衡量移动性的方法，被证实是改善疾病传播现有模型的一个重要方法。本章关注能够使航空和航海数据更精确的方法，以便得出关于一些特定全球疾病载体的更精确的观点，比如严重的急性呼吸综合征。

模型不仅仅可以通过加入飞机和船舶的移动来进一步改善，还可以加入个人选择的不同路径。全球有超过65亿活跃手机订阅，CDRs数据可以为传统流行病学家的模型带来巨大的价值增值。另外，当CDRs数据与用户当前健康状况手机实时调研数据结合起来，手机可以成为一个强大的流行病预警系统。

一般移动性和交流行为是一个人的思想和健康状况的很好的指标，但搜索词可以提供额外的信息：它们表明了某人关于某个话题的明确兴趣。正如上一章中讨论的，谷歌全球搜索词的资料库可以提供一段时间内搜索词之间的定性比较，正如类似谷歌趋势之类的工具所

展示的。正是有特定的应用目标，比如监控人们对流感的关注度，搜索词才可以成为公共健康领域独特而强大的数据库。谷歌流感追踪系统是一个使用现有大数据的优秀例子，它可以智能分析搜索词并将其与现有的流感趋势数据库配对，让任何人都可以近乎实时地看到流感传播状况。谷歌的研究人员利用这一概念并将其扩展应用到登革热上，说明了它在追踪季节性流感以外应用的可能性。

和谷歌类似，脸谱网和推特拥有关于人们的意图、想法和当前状态的海量数据。近年来，特别是推特，已经成为公共健康研究人员的目标之一。研究人员挖掘在推文中表达的健康观点，并开发自然语言处理运算法则来消除推文之间的歧义（比如，描述个人的症状与关于某种流行病更广泛传播的新闻）。这个研究的早期阶段主要关注流感追踪，因为之前就有的数据库可以用于验证它们。但一些研究人员扩展了研究范围，开始检索更加综合的个人健康状况描述，比如过敏、肥胖以及其他病痛等。

本章强调了全球数据对于公共健康可能的影响，我们希望可以在流行病肆虐之前识别并且能消除它们。

航空线路数据和疫病传播

世界各国长期以来都在记录哪些人患了哪些疾病。国家数据库记录这些数据已经有数十年的历史了，比如在英格兰和威尔士，从1948年开始到现在都有关于麻疹传染的双周记录数据。这些历史数据库对于数学家和流行病学家来说很有吸引力，因为它们可以用于创建疾病传播的数学模型，利用数据库可以识别疾病的起源并量化它的移动。数学家将这些可以描述任意一种在某个系统内的传播模式的已有传播模型，与历史数据模式相比较。通过校准模型中的参数，他们试图找

到一个可以复制这些数据的模型，其目的是找到一个最匹配的模型以用于预测未来疾病的传播。

这些模型通常包含一个描述人口流动的参数，该参数是确定传染性疾病传播方式的一个重要因素。历史上曾经通过假设人们随机移动的方式来预测人类的流动性，还曾通过相当粗略的调研来预估，即询问成数百至数千人的小样本人群在过去一周、一个月或一年中他们都是怎么移动的。但这种类型的调研有许多先天缺陷：不准确（依靠人们错误的记忆）、样本偏少（每次只能调研一小部分人）且时效性有限（传统调研并不是持续进行的）。

然而，在21世纪早期，流行病学家和数学家意识到，他们可以挖掘出能够更能代表人类在国家内或区域内的移动性的数据。前一章中讨论过的航空和海运线路，都可以用于了解人们和货物（两者都可能携带疾病载体）在全世界移动的情况。

航空线路提供了一个展示城市之间连接方式的嵌入式网络。使用往来机场的人流数据，流行病学家可以在全球疾病传播模型中增加一个信息层，这在简单地通过稀少的数据来假设人口移动性的时期是不可能的。

2006年，维多利亚·科利扎和她的同事证明了，航空运输网络在严重急性呼吸综合征等新发传染性疾病的全球传播中扮演了重要角色。研究人员甚至提供了一个基于航线网络的模型，他们宣称这一模型可以预测新发传染疾病的暴发。尽管他们也承认，使用季节变化以及区域卫生状况和卫生设施差异等其他要素，可以让模型更精确。

但并非所有的航空和海运线路在全球疾病传播中都有相同的重要性。实际上，在描述疾病传播的过程中，只有相当小的一部分城市间的航空线路会被使用到。格奥尔基·波巴谢夫、罗伯特·J·莫里斯和D·迈

克尔·戈德克认为，在3 000个城市中选取200~300个之间的样本，就可以获得关于疾病传播的足够信息。

对有些全球性的疾病来说，如通过蚊子或其他害虫传播的疾病，可以在航空乘客路线之外加入全球海运路线来衡量移动性从而更好地建模，因为大型货船运输的货物很容易成为昆虫搭乘的便车。比如，疟疾是通过不同种类的蚊子传播，冈比亚按蚊是传染疟疾的一种主要物种，20世纪30年代它们从非洲传到巴西就是一个典型的主要通过海运路线传播的案例。近年来，海运集装箱把亚洲虎蚊，又叫白纹伊蚊引进到了新的区域，这种蚊子是登革热、黄热病和西尼罗热的重要载体。更好地了解蚊子或者其他通过海洋运输的疾病载体在全球的移动，可以帮助港口城市采取预防措施来防范害虫传播的疾病。

疾病预测

尽管航空和海洋数据可以改善模型，最有效的调整这些模型的方法还是关注人类的移动行为本身，而不仅仅是这些数据表现。关于人们如何移动的精细信息，可以通过CDRs数据来采集（像第七章中提到的那样）。从CDRs数据中获得的移动性信息可以反馈到流行病学模型中，以更准确地了解移动情况，因为在全球模型中不同区域的移动情况是有差异的。

根据GSMA无线情报机构的研究，到2012年年末全球有32亿人每人拥有至少一部手机，而到2017年这一数据将会达到40亿（然而，这些数据仍然远远少于真实的活跃手机数量，因为有些人有不止一部手机，2012年年末的活跃手机数量大约是60亿部）。这些活跃的手机产生了拍字节级别的数据，在全球每个国家高效地留下了人们的位置和交流的数字踪迹，而且这些数据可以近乎实时地被获取到。

CDRs的历史数据对人类的流动性提供了实证的、精细的观察，可以用于改善模型，就像航空和海运数据一样。CDRs数据包含了移动状况的季节性差异和区域性差异。尽管多数流行病模型假设了更加静态和粗放的条件，但采用了CDRs数据的模型可以更容易适应不同的时间和区域。

但重要的是，CDRs数据本身也可以作为疾病暴发的一种新的指标，一个可以实时更新的指标。为了证明这一观点，马克·利普斯奇和他的同事在2010年甲型H1N1流感暴发时期，在墨西哥部署了一个面向100万手机用户的移动平台。与第九章中提到的Jana公司案例类似，该平台让用户参与关于他们当前健康状况的调研以换取一定量的通话时间。当调研结果和通过CDRs数据获取的被动信息结合在一起的时候，可以发现手机用户的移动和交流模式的变化与疾病的暴发呈现可能的相关性。

传统的流感区域预警系统最少也有几天乃至几周的延迟，因为有症状的人不会马上去看医生，当地政府和医院也不会马上把诊断案例更新到数据库中。一个能够自动挖掘实时CDRs数据和手机调研数据的工具，可以改革传染性疾病监测机制，这对于发达和发展中国家来说都可以实现。

新的监测机制的实施并不需要花费太多。流行病学家和政府部门可以直接利用现有的手机基础设施及其采集的数据，建立一个真正的预警系统，而不需要去更新、修改或是扩展现有的疾病报告系统。他们还可以挖掘其他数据来源，比如橙汁销量或者火车站所拍摄到的乘客咳嗽的监控视频。这一迅速发展的领域被称为“症状监测”，其本质是通过一系列分散的多样数据来源，在早期阶段发现集聚的疾病症状。政策制定者通过使用这些数据的系统可以建议可行的早期行动，比如关闭中学和大学，及时避开疾病可能带来的严重影响。

用数百万人的网页搜索预测感冒的活跃度

2009年，《自然》杂志发表的一篇论文揭开了大数据的秘密。该论文由谷歌的研究人员撰写，展示了如何利用数百万人的网页搜索数据，来推测一天中流行性感冒在美国的不同区域的活跃度。它说明谷歌的搜索词条数据库除了用于呈现受欢迎的网页或有关的广告之外，还可以发挥更大的作用。谷歌找到了利用用户的兴趣来推断他们的健康状况的高效方法，并将该方法运用到了设计谷歌流感追踪系统这一免费工具。

谷歌流感追踪系统是在数千亿条美国搜索数据的基础上开发出来的，包括2003~2008年这5年时间的匿名谷歌网页搜索记录。但为了找到与追踪流感周期关系最密切的搜索词，研究人员研究了美国疾病控制与预防中心的流感预防定点监控网络的公共数据。他们基于美国疾病控制与预防中心的公共数据建立了一个流感活跃度模型，然后和他们未经过滤的5 000万条最常用的搜索词条数据库对比。美国疾病控制与预防中心模型和搜索词条数据库之间的对比显示，在频率和时间方面而言，特定的搜索词条出现的高峰和美国疾病控制与预防中心数据显示的流感活跃性的增强相一致。通过对比，研究人员获得了与美国疾病控制与预防中心模型中每年流感活跃性相符的45个搜索词条，并将这些词条归入流感并发症、感冒/流感药品、常见流感症状以及其他与流感相关的大类中。

这个验证搜索词与流感活跃性相关的方法，被重复用于至少28个国家以及美国的全部50个州。不同国家流感活跃性的实证数据库可以在谷歌.org网站上查看。此外，流感追踪数据可以生成随时间变化的活跃度图表，也可以以文本文件格式下载。

在一个称为登革热追踪系统（Dengue Tracker）的工具中，谷歌把确定流感活跃度的方法扩展到了登革热的研究上，并在包括玻利维

亚、巴西、印度、印度尼西亚和新加坡在内的至少十个国家应用。2011年的一篇论文中介绍了研究人员开发该工具的方法，并提到这个工具可能具备特别大的用途，因为登革热暴发的国家大多缺乏传统疾病监控资源。

但它的算法并不是完美的，而是其他大数据模型一样，仅仅基于搜索词的模型需要核查。谷歌的数据认为2013年1月在流感季暴发的时候有11%的人口感染了流感，但这一预测数据几乎是美国疾病控制与预防中心所监测到的6%的两倍。一些研究人员猜想，这个差异可以用流感季大量的新闻报道引发了社交媒体上的大讨论来解释，但究竟是什么因素仍然不清楚。这表明了对于模型来说，如果采用了一定量的搜索信息，那么最好也能包含一些实证数据的现实核查，我们将在下一节中继续讨论这一点。

流行病网络

推特和脸谱网可能是世界上拥有最明确、最详细的个人信息的两个在线服务商。人们所公开的关于自己的信息总量是令人难以置信的，从工作、位置、政治立场乃至状态更新中透露的个人观点无所不有。迄今为止，很少有研究人员可以从脸谱网数据中挖掘全球公共健康的线索。相反，近年来很多研究人员已经开始利用相对比较好获取的推特数据。他们的初期成果展示了利用社交数据建立综合可靠的公共卫生系统的光明未来。

2010年，瓦西里欧斯·兰波斯和内罗·克里斯蒂安尼尼扩展了谷歌研究人员的工作，他们通过挖掘英国550万活跃推特用户每天发布的数十万条推文，来寻找流感的相关指标。和谷歌研究人员用来寻找与区域内流感发生率相关的搜索词的方法相似，兰波斯和克里斯蒂安尼尼比较了2009年6~12月间英国已有的流感相关数据。研究人员发现，特定

的含有流感指标的推文，与英国健康保护局2009年甲型H1N1流感暴发期间数据的相关性超过95%。基于推特的流感追踪系统是对谷歌研究方法的一个独立验证，并且提供了改善单一使用搜索词的可能。尽管在线搜索和推文都可能会被媒体的大肆宣扬和讨论所影响，但人们在推特环境下可能会提供更明确的背景信息，比如发表类似“我感冒了”的推文。如果自动系统检测到这类推文，就可以用于推测该情形在普通人群中的规模，这与第九章中提到的手机调研的方法类似。

其他研究人员，包括荒牧英治和他的同事，以及辛西娅·周还有巩特尔·埃森巴赫，也用不同的方法分析了推文。荒牧开发了自然语言过滤器来区分有关流感的讨论和表明一个人生病的推文。周和埃森巴赫关注了三种方法：监控诸如H1N1这样的术语词汇，使用内容分析来确定推文的含义（比如判断是分享新闻还是自我诊断），以及验证推特作为实时追踪工具的效果。

另一些研究人员将推特看作一个全球层面更综合的公共健康数据来源。迈克尔·J·保罗和马克·德雷泽开发了一个主题模型，是特别为寻找描述症状和不同常见疾病的治疗方法的词汇而设计的。常见疾病话题模型通过对包含疾病讨论的160万条推文进行学习，已经可以提取从流感、传染病到受伤、牙齿问题、一般疼痛等状况的信息。研究人员宣称，该模型与谷歌追踪系统的早期追踪结果以及其他建立在政府健康数据库基础上的推特模型相符。

推特数据挖掘展示了一种采集公共健康观点的可能的新方法，从某种意义来说，就是监听公众对于健康的讨论，而不是直接询问人们的看法。尽管如此，在一个社会健康监控系统完善之前，还需要做很多工作来验证来自大家的观点，也许需要跟手机调研结合起来。

作为一个全球社区，我们处在公共健康新时代的突破点上。长久以来，人类都受到传染性疾病的困扰，但随着医学的发展以及科学家对疾病原因的研究进展，很多生命得到拯救，人类承受的病痛也被缓

解。然而，尽管现代科学家开始了解疾病的来源，但他们并不能一直清楚地掌握疾病在人群中的传播途径，也不能肯定地预测疾病传播的精确路径。

如今，科学家终于看到了终极的预防性工具到来的曙光——大数据。大数据和现实挖掘有可能可以创造一个实时“水晶球”，可以及时将新型流感病毒或者霍乱发病率的突然上升这类紧急疾病动态信息，及时通知医疗服务人员、公共政策制定者和其他人。

第五部分重点提到了一些可以为这个“水晶球”添砖加瓦的数据来源，从出行网络和搜索引擎检索词到手机调研和动态推文，同时还着重介绍了一些使用这些数据的项目。随着时间的推移，急性流行病的最佳预警系统可能会是建立在许多数据来源的组合上的，并且这些数据可以回溯到个人以进行现实核查（通过移动设备调研或通过能核实的推文）。同时，我们还需要更多的工作，来找到通过社交网络、搜索词条以及手机数据推断关于慢性疾病的最佳方法。我们也需要更好地了解我们的数据怎样才能和慢性疾病联系起来，比如糖尿病、免疫系统紊乱以及心脏疾病。正如流感数据在谷歌工程师的眼皮下被埋藏了将近十年的时间，慢性疾病的相关数据也可能就近在眼前，它们只是需要从噪声里被筛选出来。

结语

在本书的写作过程中，我们决定尽可能多地关注那些进行现实挖掘的初创企业或知名公司。尽管也有很多有趣的学术论文，但是它们往往关注一些短期研究项目，这些项目不一定能生产有深远影响的计划，或是只能生产难以验证的一次性成果。当然，这些公司可能会昙花一现、被兼并或是消失，但我们相信，分享已有的这些现实挖掘案例，可以更好地在实践领域夯实大数据应用的基础。

也就是说，我们在编辑过程中，逐渐删掉了本书初稿中提到的一些公司，因为它们已经不存在了。并且，在本书付梓出版的时候，还会有一些公司可能在未来3~5年中消失。大数据的世界在飞快地运转，而这些日新月异的初创企业正是这一速度的明证。

但是，我们也认为学术论文在大数据领域占有重要的一席之地。毕竟“现实挖掘”这一术语就是在麻省理工学院的一篇研究论文中首先提出的。本书中引用的论文只是探索了大数据应用的一些基本的可能性。而真实的情况是，为了实现海量数据的全部潜力，并给人类系统提供周到而缜密的应用程序，软件工程师们需要面对和处理的数据量十分巨大，无论是范围层面还是时间规模，都是学术研究中几乎不太可能实现的。

此外，作为一本大数据挖掘指南，本书的核心是号召所有人行动起来。如果你是企业家，请想想通过现实挖掘可以为你、你的邻居和世界提供些什么；如果你身处政府部门，请思考如何利用数据来制定更好的政策；如果你是科研工作者，可以考虑如何将研究项目推广到更加广泛和长期的应用领域。而在所有这些行动中，都应注重隐私问

题以及它在不同范围层面和不同社会背景下的转变。在设计解决方案的初始阶段就应该考虑隐私问题，数据的采集和使用过程也应该保持透明。

现实挖掘可以从那些失效的系统开始着手，如慢性疾病管理混乱、社区衰败、组织冗余、公路堵塞、经济衰退以及全球性传染病等。

接下来，考虑与这些系统失效有关的或可以作为指标的数据类型：生理行为的变化、街头涂鸦的增多、生产效率的下降、汽车移动速度变慢、购物习惯的转变、人们的旅行模式等。本书提供了多个相关数据集的一些指标，但必须承认样本有限。数据无处不在，使用它们只是接触数据的一种方式而已。

还需要考虑在不同层面采集数据——个人、社区和组织、城市、国家以及全球层面。这些层面上分别有什么样的隐私性问题需要考虑？有些什么样的数据分享激励措施？透明度在何时何处最能发挥作用？哪些人可以从你的数据中获益？为什么？

最后，除了从数据中获取信息外，人们还应思考如何通过数据让城市系统运行得更好更智慧。如何设计一个帮助糖尿病病人更好地监测自身状况的手机应用？如何帮助市民参与振兴衰败的社区？如何让脑力劳动者更方便地分享信息？当驾驶者有需求的时候，怎样预测交通并将预计的路程时间和备选路线等信息发送给驾驶者？如何根据消费者的特定消费模式预见到国家的经济衰退？如何帮助政府更高效地配置经济激励资金？如何根据人员移动抑制下一轮大规模流行病的暴发？

然而，如果认为我们持续产生的这些大数据只会被用于推动世界进步，那就太天真了。近期的一些事件显示，政府在获取大数据时有可能会滥用这些数据：监视居民、镇压不同政见者或是妨害公民自

由。此外，公司和营销者希望通过确定消费者的行为来获取更大利润，或是推送更有针对性的广告，或是影响消费行为，这对绝大多数人来说也不会是进步。关于这些大数据使用不当的麻烦后果，我们留待其他地方再做深入讨论。

但是，如果因为大数据可能被用于不道德应用就停止数据采集，那也是幼稚的。这就是为什么我们给工程师、企业家、学者和政策制定者提供了另一条路径：利用数据来推动积极的变化，并在使用数据的整个过程中立足现状并考虑处理个人数据时的道德约束。

我们希望本书对大数据的积极潜力进行了基本阐述，也希望通过此书让大家了数据挖掘方法的应用、系统和概念。大数据的时代就在眼前，让我们一起来建造一个更美好的世界吧！